Vision-and-Language Navigation: Literature Review

Jing Gu
University of California, Santa Cruz
jgu110@ucsc.edu

Qi Wu
The University of Adelaide
qi.wu01@adelaide.edu.au

Xin Eric Wang
University of California, Santa Cruz
email@domain

Abstract

Vision-and-Language Navigation (VLN) has attracted widespread attention due to its practical significance. A robot that can understand natural language instruction and interact with the environment to achieve the target goal is revolutionary for human society. This paper presents a thorough review of VLN by summarizing related papers in a structured manner. We first categorize and introduce current VLN benchmarks. Then we introduce solutions for VLN. Next, we introduce current evaluation metrics and tasks analysis. Finally, we discuss future directions.

1 Introduction

Recent advances in deep learning, natural language processing, computer vision, and reinforcement learning have boosted the development of embodied AI. Recently, Vision-and-Language Navigation (VLN) has been getting more attention since the proposal of Room-to-Room dataset (Anderson et al., 2018b). Vision-language navigation (VLN) is the task of navigating an embodied agent to carry out natural language instructions inside real 3D environments.

Researchers have achieved success on Visual Navigation in both simulated environments (Zhu et al., 2017; Mirowski, 2019) and real environments (Mirowski et al., 2018). But it does not involve text modality. Artificial intelligent agent in VLN is capable of understanding natural language instruction or even asking helps during the navigation process, which has a wider applicable scenario in human society.

Unlike existing natural language robots control tasks, VLN requires agents to execute instruction in an unseen environment with photo-realistic scenes. Over-simplified 3D environment structures inevitably leads to agents' poor performance in real life. By using real images rather than rendered ones (Beattie et al., 2016; Kempka et al., 2016; Zhu et al., 2017), the agents build in VLN environment can be better adapted to real life scenarios. Another major difference is that VLN agents might actively seek more information through natural language rather only passively receive a short command from human or another agent. More powerful language ability enable an navigable agent to achieve more complicated tasks.

Since the development of Room-to-Room (Anderson et al., 2018b) dataset, VLN has witnessed tremendous development in both benchmarks and solutions. Environment has extended from indoor such as apartment to outdoor such as street. The instruction form extends from detailed route instruction to high-level goal instruction. The agent is required to understand information from textual modality and visual modality to learn the environment, and to navigate with sophisticate strategy to reach the target goal. Each path in VLN contains complex instructions and a sequence of changing environment images. Various data-centric methods are proposed. Methods are also proposed to learn a better representation of the environment and the instruction. The outcome of a navigation process depends the navigation action decision in each step. Methods are also proposed to learn a better action strategy.

We structure this paper as follows. In section 2, we categorize VLN tasks and introduce them separately. In section 3 we talk about the current methods used for building VLN agent. In section 4, we introduce commonly used evaluation metrics and VLN tasks analysis. Finally, we discuss future directions in section 5.

2 Tasks and Datasets

Many VLN tasks have been proposed out of the significant value in real life. Vision-and language navigation has been an increasingly popular area. In order to build appropriate testbeds for the newly proposed task, various benchmarks have been developed. The visual features of the datasets are deeply connected with the simulators. Please refer to Appendix A for more details about the datasets and refer to Appendix B for more details about simulator commonly used to create VLN benchmarks.

Here we categorize VLN benchmarks according to two dimensions, object availability level and language level. object availability level defines how hard it is to find the target goal. We set three levels. In the first level, the agent can find the target according to detailed step-by-step route description. In the second level, the agent is required to find a remote target goal with a coarse navigation description, which means the agent needs to reason a path in an unseen environment itself. In the third level, besides reasoning a path, the agent also need to manipulate the environment objects to achieve the goal since the object might be hidden or need to change physical status. Another dimension language ability level describes how the agent is capable of natural language understanding and expression. We also set three levels here. In the first level, the agent can understand a natural language instruction at the beginning. In the second level, the agent is capable of sending signal for help whenever it is unsure, and then understand additional instructions. In the third level, more like human, the agent asks questions in the form of natural language during the navigation, and understand guidance for next steps in order to the navigation.

2.1 Route-Detailed Navigation Tasks

Given a detailed natural language instruction, an agent needs to strictly follow it to reach the goal target. Anderson et al. (2018b) built Matterport3D simulator and further created Room-to-Room (R2R) Navigation benchmark based on Matterport3D panoramic RGB-D dataset (Chang et al., 2017). An embodied agent in R2R moves through a house in the simulator jumping between adjacent pre-defined panoramic viewpoints. Amazon Turker manual wrote the instruction for each given path in the environment. Each path has three different instructions. The instruction will be provided at the beginning of the navigation process to provided a detailed moving path. Here is an example: ”Head upstairs and walk past the piano through an archway directly in front. Turn right when the hallway ends at pictures and table. Wait by the moose antlers hanging on the wall.” Check Appendix D for the leaderboard of R2R dataset.

All navigation paths in R2R contain at most six edges and are shortest-to-goal paths. Therefore, reaching the goal destination is not strongly connected with following the language instruction. Jain et al. (2019) joins one path with another path whose start point is the endpoint of the former one. Likewise, Room-6-Room (R6R) and Room-8-Room (R8R) (Zhu et al., 2020b) are further proposed to generate VLN dataset with longer paths.

R2R is also extended to includes languages other than English. Ku et al. (2020) proposed Room-Across-Room (RxR). RxR contains instructions from English, Hindi, Telegu. The dataset has more samples and the instructions in it are time-aligned to the virtual poses of the instruction. Yan et al. (2020) collected Room-to-Room (XL-R2R) to extend R2R with Chinese instruction.

In most current benchmarks, agents navigate through pre-defined viewpoints. Krantz et al. (2020) reconstructed the nav-graph-based R2R trajectories in continuous environments and created Vision-and-Language Navigation in Continuous Environments (VLNCE). Irshad et al. (2021) proposed Robo-VLN task where the agent has continuous action spaces over long-horizon trajectories.

Various outdoor benchmarks are also proposed. Mirowski et al. (2019) introduced the StreetLearn task for outdoor navigation with photographic content from Google Street View. Mehta et al. (2020) release raw Street panoramas need for TOUCHDOWN as an addition to the StreetLearn dataset. StreetNav extends StreetLearn (Hermann et al., 2020) with additional driving instructions from Google Maps by randomly sampling start and goal positions.

Based on Google Street View, Chen et al. (2019) introduced the TOUCHDOWN. In this outdoor VLN task, an agent first follows navigation instructions with real-life observations in New York City and then identifies a location through reasoning in natural language to find a hidden object.

Talk to Nav (Talk2Nav) (Vasudevan et al., 2021) is an interactive visual navigation environment. The dataset is annotated by Amazon Turker and the navigational instructions are more realistic.

LANI (Misra et al., 2018) is a 3D navigation environment and corpus, where an agent navigates between landmarks following natural language instruction. They first generate reference paths that pass near landmarks, then use Amazon Mechanical Turk to annotate.

Blukis et al. (2018) proposed a quadcopter flying navigation task based on randomly generated virtual environments in Unreal Engine. Following the setting of LANI (Misra et al., 2018), Blukis et al. (2019) create a quadcopter drone navigation task, where the navigator flies between landmarks following natural language instruction.

The pre-defined navigation instruction sometimes is still ambiguous to the complexity of the environment. In that case, help from oracle would be necessary. Chi et al. (2020) introduced Just Ask, a task where an agent could ask humans for help during the navigation process based on the Room-to-Room dataset. To simulate a human who could also occasionally makes mistakes, the oracle agent who has access to the shortest path information could give incorrect answers with a certain probability. For now, no datasets requires agent to actively ask more information in natural language after given detailed route instruction. We do think it is necessary if the environment is challenging that the detailed initial instruction is not sufficient.

2.2 Object-Targeting Navigation Tasks

In real life, route-detailed instruction could even be impossible such as in an unseen environment. More often, instructions are more concise and contain merely information the target goal, such as ”Bring a spoon”.

Wu et al. (2018) proposed Concept-Driven Navigation (RoomNav) task based on House3D environment. The goal is in the form ”go to X”, where X is a pre-defined room type or object type. Qi et al. (2020b) proposed Remote Embodied Visual referring Expression in Real Indoor Environments (REVERIE), in which given a remote object, the agent locates the target with high-level natural language instructions. The agent navigates and finds the object from multiple distracting candidates. Zhu et al. (2021a) proposed a object location task SOON where the agent is limited in a given path during the searching process.

Based on Mattherport3D dataset, Nguyen et al. (2019) proposed Vision-based Navigation with Language-based Assistance (VNLA), where an agent is trained to navigate indoors to find the target objects by requesting and understanding the language instructions from humans. Hanna (Nguyen and Daumé III, 2019) is introduced to build a navigation agent that could utilize help from language assistants based on Matterport3D simulator. An agent solving the tasks needs to request next-step instruction CEREALBAR (Suhr et al., 2019) is a collaborative task between a leader and a follower. The two agents move in a virtual game environment to collect valid sets of cards. The follower only has a first-person view and executes the instruction from the leader who has access to the whole map.

In reality, navigation robots may use language to ask for assistance and take actions based on the feedback. Cooperative Vision-and-Dialog Navigation (CVDN) (Thomason et al., 2019b) is a dataset of human-human dialogs. Besides deciding on the next action, the navigation agent also needs to ask questions for guidance. The oracle with extra information about the next best steps needs to provide answers. Thomason et al. (2019b) also proposed a simplified version, Navigation from Dialog History (NDH) task, in which an agent will be trained to navigate without asking questions and further utilizing the response from the oracle. Banerjee et al. (2020) presented Localization and Mapping with Natural Language (RobotSlang) for cooperative robot navigation. Dialog could also be helpful in complex outdoor environment. de Vries et al. (2018) introduced Talk the Walk dataset, in which a guiding agent and a tourist agent interact with each other to have the tourist navigate towards the correct location. The guide has access to the map but does not know where the tourist is. The tourist navigates a 2D grid via discrete actions.

2.3 Manipulation-added Navigation

The target object might be hidden (The spoon is in a draw), or the target object needs to be manipulated (Need an sliced apple but only uncut one found). In these scenarios, manipulation on objects is necessary to accomplish the task. Based on indoor scenes in AI2-THOR, Shridhar et al. (2020) proposed ALFRED dataset, in which agents complete household tasks in an interactive visual environment. TEACh (Padmakumar et al., 2021) is a dataset that studies object interaction and navigation with free-form dialog.

Task Complexity Interation w/ Oracle	Route-Detailed	Object-Targeting	Navigation + Manipulation

No Interaction	*Matterport3D group;
	*Google Street View Group;
	LANI (Misra et al., 2018)
		REVERIE (Qi et al., 2020b); RoomNav (Wu et al., 2018); SOON (Zhu et al., 2021a)	ALFRED (Shridhar et al., 2020)

Guidance Only	Just Ask (Chi et al., 2020)	VNLA (Nguyen et al., 2019); HANNA (Nguyen and Daumé III, 2019); CEREALBAR (Suhr et al., 2019)	-
Dialog	-	CVDN (Thomason et al., 2019b); RobotSlang (Banerjee et al., 2020); Talk the Walk (de Vries et al., 2018)	TEACh (Padmakumar et al., 2021)

Table 1: Vision-and-Language Navigation datasets. *Matterport3D group includes Room-to-Room (Anderson et al., 2018b), Room-for-Room (Jain et al., 2019), R6R, R8R (Zhu et al., 2020b), Room-Across-Room (Ku et al., 2020), XL-R2R (Yan et al., 2020), VLNCE (Krantz et al., 2020). *Google Street View includes StreetLearn (Mirowski et al., 2019); StreetNav (Hermann et al., 2020), TOUCHDOWN (Chen et al., 2019) Talk2Nav (Vasudevan et al., 2021).

3 Methods

Many methods are proposed from different perspectives for VLN tasks. In the original setting (Anderson et al., 2018b), the agent can not explore the test environment before executing the VLN tasks. Whether have prior access to explore the test environment or not are under different VLN setting, and we firstly introduce the prior exploration methods.

Based on the given training data and optional additional data, the VLN agents understand the vision and language information and make navigation decision each step to reach the goal target. For the methods that could be used in both settings, we categories them into Representation Learning, Action Strategy Learning, Data-Centric Learning.

3.1 Prior Exploration in Test Environment

There is a performance gap between the seen environments and unseen environments for VLN agents. In the original setting of VLN (Anderson et al., 2018b), the agent should be tested in an unseen environment. Meanwhile, allowing an agent to freely explore the new environment before executing tasks is also of practical benefit because it helps to adapt to a new environment such as house or apartment. Note the performance in this setting should not be directly compared with the one without prior exploration for agent evaluation purpose. Here we survey the methods for prior exploration in Test Environment.

(Fried et al., 2018) explores various possible trajectories following the instruction in each session. Then they rank the exploration trajectories based on the likelihood of the instructure based on the prediction out a trained speaker model. The exploration process could also offer more fine-tuning signal to further update the VLN model. Wang et al. (2019) first introduced a matching critic measuring the compatibility between the instruction and the navigation path. For an unseen environment where no label navigation data are available and given textual instruction, the navigator is trained to follow the path the matching critic gives a high score.

Besides exploring different possible trajectories every time before starting each session, the agent could also explore once to construct a overview about the environment. Chen et al. (2021a) proposed to build topological maps after the agent freely explored the environment. They leverages attention mechanisms to predict a navigation plan in the map and then execute the plan in action space. Zhou et al. (2021) considers the shortest-path route prior and regard VLN as a node classification problem after pre-exploring the environment.

3.2 Representation Learning

Understanding of information from both text and vision is essential for correct navigation decision. Here we introduce methods towards a better understanding of the perceived information.

3.2.1 Pre-training

Due to the excellent performance of the transformer-based pre-training model in various areas, it is natural to apply pre-training technology into the VLN model. Pre-training provides a relative good initialization from the knowledge in a close domain.

To take advantage of the natural language understanding ability of the pre-training model, Li et al. (2019) proposed PRESS to leverage large pre-training model BERT (Devlin et al., 2019) and GPT (Radford et al., 2019) as the encoder to improve the instruction understanding and robustness on the unseen environment. Pashevich et al. (2021) proposed Episodic Transformer (E.T.) that is equipped with various encoders to process different modalities. They pre-train the language encode by predicting synthetic instruction.

Visual information process also proves beneficial. Hao et al. (2020) trains a visual-textual BERT model on a large amount of image-text-action triplets from scratch to learn textual representations on VLN tasks. Their proposed PREVALENT takes image-text-action triplets as input and is trained to predict the masked tokens and the next action in a self-learning paradigm. Qi et al. (2021) proposed an an object-and-room informed sequential BERT to encode instructions and visual observations in words and object level, which strenghthen between the objects' mention and objects' visual features.

A closer relation between the pre-training task and down-stream task usually leads to a better performance. Researchers also explored pre-training on VLN task directly. By viewing VLN as a path selection problem, VLN-BERT (Majumdar et al., 2020) pre-trains navigation model from web image-text pairs to measure the compatibility between path and instruction based on ViLBERT (Lu et al., 2019). VLN can be considered a partially observable Markov decision process, where the future rendered scenes are dependent on the current scene and navigation action. Meanwhile, the transformer-based pre-training model usually does not consider such interaction with the environment by design. Hong et al. (2021) equips a BERT pre-trained on visual-textual knowledge with a recurrent function. A BERT-structured model leverages the history representations without increasing module volume by re-using the state from the CLS token.

Researchers also collected extra pre-training corpus for VLN. Guhur et al. (2021) collected a large-scale in-domain dataset for pre-training on vision-and-language navigation task. They further built Airbert based on the proposed dataset and achieved good performance on few-shot setting. Majumdar et al. (2020) also proposed to use the resourceful internet contains potentially large-scale pre-training material.

3.2.2 Graph Representation

Objects in visual scenes and concepts in instructions has a strong logical relation. Building a graph for the relation offers a better a explicit representation for these information. Hong et al. (2020a) proposed Language and Visual Entity Relationship Graph for modelling the inter-modal relationships between text and vision, and the intra-modal relationships among visual entities to capture and utilize the relationships. They used a message passing algorithm for propagating information between language elements and visual entities in the graph which determines the next action to take. Building a graph to represent the environment also provides valuable guidance. Deng et al. (2020) introduced Evolving Graphical Planner (EGP) that constructs a graph of the navigable environment during the exploration process. The navigator reasons over the graph to select the next navigable node among action space. Anderson et al. (2019) proposed a mapper that constructs a semantic spatial map on-the-fly during navigation, and an end-to-end differentiable Bayes filter to identify the goal by predicting the most likely trajectory through the map according to the instructions.

3.2.3 Memory Structure

VLN could involve prolonged navigation steps (Padmakumar et al., 2021; Krantz et al., 2020; Zhu et al., 2020b). Effetively remembering and utilzing important information during the navigation history is essential for making correct decision. Toward this, various specific memory structures are proposed. Lin et al. (2021) introduced a Memory-augmented attentive action decoder to help the agent to learn where to stop and what to attend to. Nguyen and Daumé III (2019) proposed a memory-augmented neural agent to model decision making at a different level and an imitation learning algorithm that teaches the agent to avoid repeating past mistakes. Zhu et al. (2020c) proposed Cross-modal Memory Network (CMN) to remember relevant information in both textual modality and image modality. To effectively memorize information in long trajectory, Chen et al. (2021b) proposed a hierarchical encoding of the panoramic observation history

3.2.4 Attention Structure

Concepts and objects that are relevant to current step decision making should receive a high attention. Various attention structures are also proposed toward a better information understanding of textual and visual modality.

Landi et al. (2019) employed dynamic convolutional filters to attend the visual scene and control the actions of the agent. The convolutional filters are produced via an attention mechanism and are also utilized in the scene to which the navigator moves. (Zhang et al., 2020a) designed a cross-modal grounding module composed of two complementary attention mechanisms to track the correspondence between the textual and visual modalities. Landi et al. (2020) presented Perceive, Transform, and Act (PTA), where text, vision, and action are merged with a fully transformer-based model without the usage of the recurrent function. (Qi et al., 2020a) proposed Object-and-Action Aware Model (OAAM) that processes object description and direction description contained in the instruction separately. Gao et al. (2021) proposed Room and-Object Aware Attention (ROAA) mechanism to explicitly perceive the room- and object-type information from both instruction and visual observations.

3.2.5 Multi-task Learning

VLN shares similar modality understanding with other Vision-and-Language tasks and thus it is possible to improve both via multi-task learning.

Visual representation model tends to overfit in seen environments. Wang et al. (2020c) proposed an environment agnostic learning method to learn a visual representation to generalize better on unseen environments. The learning target is to build a latent representation that is invariant among the seen environment to mitigate the overfitting problem. Chaplot et al. (2020) proposed attention module to train a multi-task navigation agent to follow instructions and answer questions.

3.2.6 Auxiliary Tasks

A VLN environment and instruction usually contain much information that is relevant to the task goal. Therefore, except only using the signal from the navigation target, various auxiliary tasks are proposed to utilize the informative environment.

Textual instructions in VLN may contain step-by-step guidance. The navigator needs to be aware of the navigation progress and the relation between the navigation path and instructions. Wang et al. (2019) train a matching model to score the degree to which the navigation path follows instruction. The matching score is used as a reinforcement learning reward signal and a direction signal in an unseen environment. Ma et al. (2019a) introduces Self-Monitoring navigation to estimate progress made towards the goal using vision and language co-grounding. The proposed agent identifies the correct direction by finding which part of the instruction is aligned with the current observation. The navigator leverages a progress monitor to estimate the distance between the current viewpoint and the final goal position, conditioning on navigation history and instruction. To better utilize the rich semantic information contained in the environments, Zhu et al. (2020a) proposed four self-supervised auxiliary reasoning tasks to help the agent to acquire knowledge of semantic representation, i.e., 1) trajectory retelling task to explain previous actions; 2) progress estimation task to evaluate the task completeness; 3) angle prediction task to predict the turn angel for next step; 4) cross-modal matching task to align the vision and the language information. Without extra labeling cost, Huang et al. (2019b) defines Cross-Modal Alignment to assess the fit between instruction and path, and Next Visual Scene, which predicts latent representations of future visual inputs in the path.

3.3 Action Strategy Learning

VLN agents navigate a long path with many decision steps to reach the target goal. With the necessity to make decision at each step, the agents faces with tremendous navigation paths choices in a complex 3D environment. Here we introduce methods on the action strategy learning process.

3.3.1 Instruction Conditioned Navigation

(Kurita and Cho, 2020) introduced a generative language-grounded policy that considers available actions at each step and use Bayes' rule to obtain the posterior disctribution over actions conditioned on the natural language instruction. The agent navigates in the environment by choosing an action that maximizes the probability of the entire instruction on a language model.

3.3.2 Exploration during Navigation

Exploring the environment and gathering information during the navigation process provides a clear observation of the local environment. Beam search is a commonly used exploration strategy for finding a better trajectory (Anderson et al., 2018b; Ma et al., 2019a,b; Tan et al., 2019). It avoids bad actions in next step by looking ahead and scoring multiple global trajectories. More sophisticated exploration strategies are also proposed toward the complexity of the VLN task.

Wang et al. (2020a) developed a more active exploration module on unseen environments. The agent learns to select among visual features of different navigable views and a STOP action. The agent stops exploring and makes navigation decisions when choosing the STOP action. They further designed a recurrent network-based multi-step exploration mechanism until sufficient information was collected.

Instead of implicitly keeping information about the environment into neural weights, Some works explicitly store the overall knowledge from exploration history. Liyiming Ke (2019) introduced Frontier Aware Search with backTracking (FAST). FAST combines global and local knowledge to compare partial trajectories of different lengths and backtrack when making mistakes—the agent backtrack when the progress estimator gives a low score for the current action. Ma et al. (2019b) proposed Regret Module based on a progress estimator (Ma et al., 2019a). They first proposed Regret Module that decides whether to continue moving forward or roll back to a previous state. Then they introduced Progress Maker that helps to navigate according to the progress score.

Koh et al. (2021) designed Pathdreamer, a visual world model for indoor environment to synthesizes high-resolution visual observation along a trajectory in future viewpoints. VLN model benefits greatly from the generated observation without actual looking ahead.

3.3.3 Reinforcement Learning

In reinforcement learning (RL), agents aim to maximize cumulative rewards. In this way, reinforcement learning could be brought into this Markov Decision Process (MDP). A navigation agent interacts with the environment during the navigation process, making a decision based on the latest environment. An increasing number of researchers have applied deep reinforcement learning to vision-and-language navigation.

Wang et al. (2019) proposed a reinforced cross-modal matching (RCM) for VLN tasks that enforces cross-modal grounding both locally and globally via deep reinforcement learning. The real-world environment has rich dynamics. Wang et al. (2018) bridge the gap between synthetic studies and real-world practices by combining model-free and model-based reinforcement learning to predict the next state and reward. One problem in implementing reinforcement learning into VLN is the sparsity of reward since it only receives success or fail signal at the end of the session. **?** proposed Soft Expert Distillation module to reward the agent higher when it takes action similar to expert's, and a Self Perceiving module that rewards according to the predicted navigation schedule. Distance to the target goal (Roman et al., 2020), matching critic (Wang et al., 2019; Ma et al., 2019a) could also provide informative reward. (Zhang et al., 2020a) proposed to recursively alternate the learning schemes of imitation and reinforcement learning to narrow the discrepancy between training and inference.

3.3.4 Communication-based Navigation

Being able to ask for help when uncertain about next action. Recently, researchers have also been building navigable agents that can send signal to request help or even communicate in natural language with humans for help.

Roman et al. (2020) proposed Recursive Mental Model (RMM) composed of a Questioner, a Navigator, and a Guide with three sequence-to-sequence models. Suhr et al. (2019) introduced a learning structure to distinguish recovery reasoning required for generating implicit actions and actions mentioned in regular instruction. Nguyen et al. (2019) proposed Imitation Learning with Indirect Intervention (I3L). An advisor modified the environment to influence the agent's decision in both training time and test time. Zhu et al. (2021c) proposed SCoA to determine whether and what to ask help from oracle.

3.3.5 Other Strategies

The agent needs to make decision in a larger action space In continuous environment. (Krantz et al., 2021) develops a waypoint prediction network (WPN) that predicts relative waypoints based on natural language instructions and panoramic vision.

3.4 Data-Centric Learning

Data is an essential component in current machine learning paradigm. Vision-and-language navigation involves information from different modalities. Methods are also proposed from different aspects of the training data.

3.4.1 Data Augmentation

Compared with the large navigation space, complicated scenes and invariance of the textual instruction, the training data is relatively sparse. Various data augmentation methods are proposed towards the scarcity of VLN datasets.

The training set could be directly augmented via generating more path-instruction pairs from the navigable environments. Fried et al. (2018) trains a speaker module to generate textual instruction given navigation paths. The proposed speaker model could further boost the performance by ranking the path candidates during test time. The generated data used for augmentation could have various levels of quality. Huang et al. (2019b) scores the cross-modal alignment to differentiate high-quality pair from low-quality pair in the data generated by Fried et al. (2018). Huang et al. (2019a) introduced a multi-modal discriminator to select valuable samples from augmented paired vision-language sequence data. They showed that a small portion of high-quality augmented data achieved similar performance with complete augmented data. Yu et al. (2020) proposed a path sampling method based on random walks to augment the training data in Room-to-Room dataset to mitigate the potential biases in the augmentation process. With the augmented data, the generalization gap between seen and unseen environments is significantly reduced. Fu et al. (2020) proposed an adversarial path sampler to generate challenging paths. They introduced an adversarial path sampler to select paths that are challenging for the navigator.

Due to the rich information contained, the navigation environment itself could also be used to generate augmented data. Tan et al. (2019) use additional training data generated via back-translation and visual features masking based on Fried et al. (2018). They randomly mask the same visual feature in different viewpoints so that the environment is unseen to the agent to some extent. Liu et al. (2021) proposed Random Environmental Mixup (REM) to split the house scenes and then to mix up to get cross-connected house scenes as augmented data. Seeing different house scenes during one navigation trip makes the agent less likely to overfit in seen houses and generalize better in unseen environments.

Textual information in VLN may also be used for data augmentation. Hong et al. (2020b) proposed Fine-Grained R2R dataset enrich Room-to-Room dataset with sub-instruction. They further introduced sub-instruction attention and shifting modules that sub-instructions at each time-step to use the alignment information between the instruction and the environment.

Data augmentation has proven essential in VLN tasks, and curriculum learning has been used to generate high-quality augmentation data. Huang et al. (2019a) used curriculum learning to train a discriminator to distinguish high-quality augmented data. They also show the efficacy of curriculum learning by extensive analysis. Fu et al. (2020) proposed adversarial path sampler (APS) that learns to select out counter-factual augmented data. The sampler keeps learning during the selection process, and thus the selected data have a closer distribution with the original dataset.

Parvaneh et al. (2020) intervenes in visual features to generate counterfactual trajectory via minimal edit. They improve agent's capabilities to generalise to new environments at test time with both training data and their counterfactuals.

3.4.2 Curriculum Learning

Curriculum learning (Bengio et al., 2009) has received widespread attention. The task's difficulty level is gradually increased during the training process as the model keeps improving learning ability.

3.4.3 Sub-Instruction

The instruction in VLN tasks provides target goals and guidance. A navigator interprets instructions in a changing environmental context. Xia et al. (2020) presented LEARN FROM EVERYONE (LEO) to leverage multiple instructions for the same trajectory. Each instruction provides a different angle to describe the trajectory, and LEO encodes all the instructions with a shared set of parameters. Instruction could be long and complicated. To better understand the relation between the visual scenes during navigation with all parts of the given instruction, Hong et al. (2020b) splits the long instruction in the R2R dataset into short sub-instructions and align the sub-instructions with visual sequences. With the enhanced sub-instructions, the agent is provided with more detailed supervision from a language perspective. They further train a sub-instruction selection model to focus on important sub-instruction.

BabyWalk (Zhu et al., 2020b) apply curriculum learning in VLN from the textual instruction perspective. Curriculum-based reinforcement learning with increasing longer instructions is leveraged to increase the language understanding ability of the VLN agent. During the training process, the instructions change from short segments of the instructions in the R2R dataset to the long instructions in the R8R dataset.

3.4.4 Extra Environment Signal

(An et al., 2021) designed multi-module Neighbor-View Enhanced Model (NvEM) to incorporate visual contexts from neighbor views since the instruction may mention landmarks out of a single view.

4 Evaluation

4.1 Evaluation Metrics

Many metrics (Anderson et al., 2018b) have been proposed to evaluate the performance of a VLN agent on the R2R dataset. 1) Navigation Error (NE) is the distance between the last node in the navigation path and the goal location. 2) Success Rate (SR) evaluates the ratio of navigation error is less than 3m. 3) Path Length (PL) is the total length of the navigation path. 4) Success weighted by Path Length (SPL) (Anderson et al., 2018a) considers both Success Rate and Path Length. 5) Oracle Navigation Error (ONE) takes the smallest distance from any node in the path rather than just the last node. 6) Oracle Success Rate (OSR) measures if any node in the path is within 3m from the target location rather than only the last one. These metrics also apply to all the R2R variants datasets as following. PC (Path Coverage) measure how well the predicted path covered the nodes on the reference path. CLS (Jain et al., 2019) is the product of the Path Coverage (PC) and LS of the agent's path with respect to reference path. It measures how closely an agent's trajectory conforms with the entire 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 captures both success and fidelity.

4.2 Task Evaluation

Agents' performance with and without prior exploration on test set should be compared directly.

Thomason et al. (2019a) Unimodal could performance very well on many multimodal tasks, including VLN. Therefore baselines should also consider using unimodal. Hu et al. (2019) had a similar conclusion that visual features could hurt the VLN models' performance. They analyzed what extent object-based representations and mixture-of experts methods can address these issues.

Jain et al. (2019) found that paths in R2R dataset are direct-to-goal shortest paths, and thus the agents does not need to strictly follow the path to reach the target goal.

(Zhang et al., 2020b) analyzed the performance difference between seen and unseen environment and observed that low-level visual features affects the agent model. They proposed to use semantic representation that contains less low-level features.

Zhu et al. (2021b) diagnose the existing VLN method on popular benchmarks. Their results show that indoor navigation agents refer to direction tokens in the instruction heavily and the agents also set the sights on objects further from the current viewpoint. They also cast doubt on the alignments in vision-and-language claimed in many models.

Generating instruction based on paths has been frequently used in data augmentation process. However, Zhao et al. (2021) found these generated instructions are on par with or only slightly better than instructions generated by a template according to human evaluation.

5 Future Directions

Data scarcity is a common problem, and various data augmentation methods have been proven helpful in general VLN tasks. However, data augmentation could be costly, especially for high-quality data. Meanwhile, there are abundant raw data on text, image, and generation navigation area. Successful adaptation into the VLN domain could lead to a tremendous performance boost.

Instead of working on a synthetic visual environment, most VLN datasets are built on photo-realistic scenes. However, there is still a gap in real-life robot navigation (Anderson et al., 2020). For example, the real robot has more action space; meanwhile, current benchmarks usually only allow agents to navigate through the pre-defined graph. To built benchmark and further build navigator in a more physical world is of practical significance.

Current VLN benchmark and methods mainly focus on tasks where only one agent is required to navigate. However, complicated tasks in real may requires the collaboration of several robots. Multi-agents VLN tasks still faces many obstacles in structure design, information communication, and performance evaluation.

The training corpus of VLN involves information from various modalities, which leads to potential hacking risk. Also, when implemented in reality, VLN agent navigates in both indoor and outdoor would have access to sensitive information such as environment scenes, human faces. Ethical concern is also not well studied in VLN.

6 Conclusion

In this paper, we presented a comprehensive survey of vision-and-language navigation. We categorized the benchmarks and the methods. We also provided the current evaluation metrics and task analysis. Finally, we discussed future direction. VLN is of significant meaning and is still not well-explored. Researchers new to this area could find this paper useful due to its systematic review.

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. 2019. Chasing ghosts: Instruction following as bayesian state tracking. In *Advances in Neural Information Processing Systems (NeurIPS)*.

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)*.

Charles Beattie, Joel Z Leibo, Denis Teplyashin, Tom Ward, Marcus Wainwright, Heinrich Küttler, Andrew Lefrancq, Simon Green, Víctor Valdés, Amir Sadik, et al. 2016. Deepmind lab. *arXiv preprint arXiv:1612.03801*.

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)*.

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*, pages 2442–2448. International Joint Conferences on Artificial Intelligence Organization. Main track.

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 *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 2459–2466.

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)*.

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)*.

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.

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)*.

Karl Moritz Hermann, Mateusz Malinowski, Piotr Mirowski, Andras Banki-Horvath, Keith Anderson, and Raia Hadsell. 2020. Learning to follow directions in street view. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 11773–11781.

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.

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*.

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.

Michał Kempka, Marek Wydmuch, Grzegorz Runc, Jakub Toczek, and Wojciech Jaśkowski. 2016. Vizdoom: A doom-based ai research platform for visual reinforcement learning. In *2016 IEEE Conference on Computational Intelligence and Games (CIG)*, pages 1–8. IEEE.

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*.

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.

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.

Yonatan Bisk Ari Holtzman Zhe Gan Jingjing Liu Jianfeng Gao Yejin Choi Siddhartha Srinivasa. Liyiming Ke, Xiujun Li. 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)*.

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 Al-Regib, Zsolt Kira, Richard Socher, and Caiming Xiong. 2019a. Self-monitoring navigation agent via auxiliary progress estimation.

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)*.

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.

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)*.

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.

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.

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)*.

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.

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.

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.

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.

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, 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. 2020c. Environment-agnostic multitask learning for natural language grounded navigation. In *European Conference on Computer Vision (ECCV'20)*.

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.

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

Felix Yu, Zhiwei Deng, Karthik Narasimhan, and Olga Russakovsky. 2020. Take the scenic route: Improving generalization in vision-and-language navigation.

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)*.

Wang Zhu, Hexiang Hu, Jiacheng Chen, Zhiwei Deng, Vihan Jain, Eugene Ie, and Fei Sha. 2020b. Baby-Walk: 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. 2021b. Diagnosing vision-and-language navigation: What really matters.

Yi Zhu, Yue Weng, Fengda Zhu, Xiaodan Liang, Qixiang Ye, Yutong Lu, and Jianbin Jiao. 2021c. 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

B Simulator

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

House3D (Wu et al., 2018) is a realistic virtual 3D environment built based on SUNCG (Song et al., 2017) dataset. An agent in the environment has access to the visual RGB signal of the first-person view, 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 from the 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 is built-in with Matterport3D (Chang et al., 2017), Gibson (Xia et al., 2018) and Replica (Straub et al., 2019) datasets.

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 manipulation, 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) House3D 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 [1] an online API that is integrated with Google Maps and 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 Methods Boundary

In this paper, we categorize methods from data perspective, i.e., data-centric, and model perspective. For model perspective, based on cognition of stages in VLN, we further divide it into two types: representation learning and action strategy learning. Although most methods fall into only one of these three categories, some methods may involve several ones. First, for some pre-training based model, it utilized a large pre-training corpus, which could be data-centric, and it also usually utilize a transformer based structure which has been proven a better memorization ability for navigation history. Also, for some methods such as VLN-BERT (Hong et al., 2021), they only contain one module which functions as both representing information and making navigation decision.

D Room-to-Room Leaderboard

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

Level	**Name**	**Simulator**	**Language-Active**	Environment
Level 1	Room-to-Room (Anderson et al., 2018b)	Matterport3D	✗	Indoor
	Room-for-Room (Jain et al., 2019)	Matterport3D	✗	Indoor
	R6R, R8R (Zhu et al., 2020b)	Matterport3D	✗	Indoor
	Room-Across-Room (Ku et al., 2020)	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
Level 2	RoomNav (Wu et al., 2018)	House3D	✗	Indoor
	REVERIE (Qi et al., 2020b)	Matterport3D	✗	Indoor
	SOON (Zhu et al., 2021a)	Matterport3D	✗	Indoor
Level 3	ALFRED (Shridhar et al., 2020)	AI2-THOR	✗	Indoor
Level 4	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
	NDH (Thomason et al., 2019b)	Matterport3D	✗	Indoor
	RobotSlang (Banerjee et al., 2020)	-	✓	Indoor
	Talk the Walk (de Vries et al., 2018)	-	✓	Outdoor
	TEACh (Padmakumar et al., 2021)	AI2-THOR	✓	Indoor

Table 2: Vision-and-Language Navigation datasets. Language-Active means the agent needs to use natural language for help.

Simulator	Photo-realistic	3D
House3D	✓	✓
Matterport3D	✓	✓
Habitat	✓	✓
AI2-THOR	✗	✓
Gibson	✓	✓
LANI	✗	✓
*Google Street View	✓	✓

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

Leader-Board (Test Unseen)	Single Run					Pre-explore					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., 2019)	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
PTA (Landi et al., 2020)	10.17	6.17	0.47	0.40	0.36	-	-	-	-	-	-	-			
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 (Liyiming Ke, 2019)	22.08	5.14	0.64	0.54	0.41	-	-	-	-	-	196.5	4.29	0.90	0.61	0.03
EGP (Deng et al., 2020)	-	5.34	0.61	0.53	0.42	-	-	-	-	-	-	-	-	-	-
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
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
DASA (Sun et al., 2021)	10.06	5.11	-	0.54	0.52	-	-	-	-	-	-	-	-	-	-
RelGraph (Hong et al., 2020a)	10.29	4.75	0.61	0.55	0.52	-	-	-	-	-	-	-	-	-	-
ORIST (Qi et al., 2021)	11.31	5.10	-	0.57	0.52	-	-	-	-	-	-	-	-	-	-
PRESS (Li et al., 2019)	10.52	4.53	0.63	0.57	0.53	-	-	-	-	-	-	-	-	-	-
PRRVALENT (Hao et al., 2020)	10.51	5.30	0.61	0.54	0.51	-	-	-	-	-	-	-	-	-	-
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	-	-	-	-	-	-	-	-	-	-
VLN-BERT (Majumdar et al., 2020)	-	-	-	-	-	-	-	-	-	-	686.62	3.09	0.99	0.73	0.01
Recurrent VLN BERT (Hong et al., 2021)	12.35	4.09	0.70	0.63	0.57	-	-	-	-	-	-	-	-	-	-
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
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

Table 4: Leaderboard of Room-to-Room benchmark as of November, 2021.