Vision-and-Language Navigation Today and Tomorrow: A Survey in the Era of Foundation Models

Anonymous authors
Paper under double-blind review

Abstract

Vision-and-Language Navigation (VLN) has gained increasing attention over recent years and many approaches have emerged to advance their development. The remarkable achievements of foundation models have shaped the challenges and proposed methods for VLN research. In this survey, we provide a top-down review that adopts a principled framework for embodied planning and reasoning, and emphasizes the current methods and future opportunities leveraging foundation models to address VLN challenges. We hope our in-depth discussions could provide valuable resources and insights: on the one hand, to document the progress and explore opportunities and potential roles for foundation models in this field, and on the other, to organize different challenges and solutions in VLN to foundation model researchers.

1 Introduction

Developing embodied agents that are capable of interacting with humans and their surrounding environments is one of the longstanding goals of Artificial Intelligence (AI) (Nguyen et al., 2021; Duan et al., 2022). These AI systems hold immense potential for real-world applications to serve as multi-functional assistants in daily life, such as household robots (Szot et al., 2021), self-driving cars (Hu et al., 2023), and personal assistants (Chu et al., 2023). One formal problem setting to advance this research direction is Vision-and-Language Navigation (VLN) (Anderson et al., 2018), a multimodal and cooperative task that requires the agent to follow human instructions, explore 3D environments, and engage in situated communications under various forms of ambiguity. Over the years, VLN has been explored in both photorealistic simulators (Chang et al., 2017; Savva et al., 2019; Xia et al., 2018) and real environments (Mirowski et al., 2018; Banerjee et al., 2021), leading to a number of benchmarks (Anderson et al., 2018; Ku et al., 2020; Krantz et al., 2020) that each presents slightly different problem formulations.

Recently, *foundation models* (Bommasani et al., 2021), ranging from early pre-trained models like BERT (Kenton & Toutanova, 2019) to contemporary large language models (LLMs) and vision-language models (VLMs) (Achiam et al., 2023; Radford et al., 2021), have exhibited exceptional abilities in multi-modal comprehension, reasoning, and cross-domain generalization. These models are pre-trained on massive data, such as text, images, audio, and video, and could further be adapted for a broad range of specific applications, including embodied AI tasks (Xu et al., 2024). Integrating these foundation models into VLN task marks a pivotal recent advancement for embodied AI research, demonstrated through significant performance improvements (Chen et al., 2021b; Wang et al., 2023h; Zhou et al., 2024a). Foundation models have also brought new opportunities to the VLN field, such as expanding the research focus from multi-modal attention learning and strategy policy learning to pre-training generic vision and language representations, hence enabling task planning, commonsense reasoning, as well as generalize to realistic environments.

Despite the recent impact of foundation models on VLN research, the previous surveys on VLN (Gu et al., 2022; Park & Kim, 2023; Wu et al., 2024) are from the pre-foundation-model era and mainly focus on the VLN benchmarks and conventional approaches, i.e., they are missing a comprehensive overview of the existing methods and opportunities leveraging foundation models to address VLN challenges. Especially with the emergence of LLMs, to the best of our knowledge, no review has yet discussed their applications in

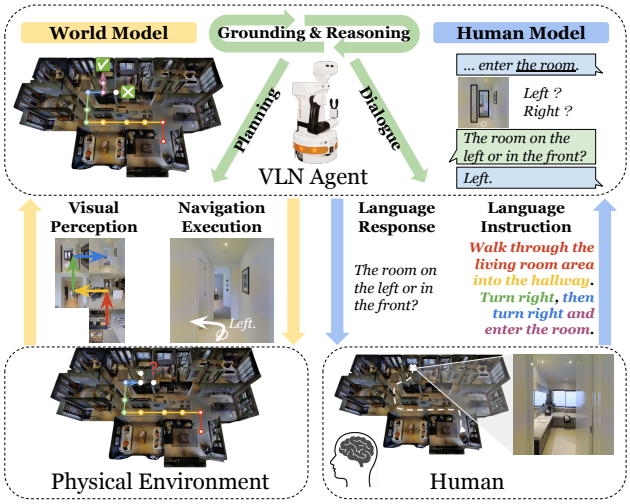

Figure 1: We organize challenges and solutions in VLN using the LAW framework (Hu & Shu, 2023). Specifically, the world model is a mental representation that encodes the visual perception and predicts the outcomes of actions. The human model is a mental representation that interprets the human intentions from textual instructions and contexts. They are a part of a VLN agent which reasons over multimodal input, engages in dialogue and plans the navigation actions accordingly.

VLN tasks. Moreover, unlike previous efforts that discuss the VLN task as an isolated downstream task, the objective of this survey is twofold: **first**, to milestone the progress and explore opportunities and potential roles for foundation models in this field; **second**, to organize different challenges and solutions in VLN to foundation model researchers within a systematic framework. To build this connection, we adopt the LAW framework (Hu & Shu, 2023), where foundation models serve as backbones of *world model* and *agent model*. This framework offers a general landscape of reasoning and planning in foundation models, and is closely scoped with the core challenges in VLN.

Specifically, at each navigation step, the AI agents perceive the visual environment, receive language instructions from humans, and reason upon their representation of the world and humans to plan actions and efficiently complete navigation tasks. As shown in Figure 1, a ***world model*** is an abstraction that agents maintain to understand the external environment around them and how their actions change the world state (Ha & Schmidhuber, 2018; Koh et al., 2021). This model is part of a broader ***agent model***, which also incorporates a ***human model*** that interprets the instructions of its human partner, thereby informing the agent's goals (Andreas, 2022; Ma et al., 2023). To review the growing body of work in VLN and to understand the milestones achieved, we adopt a top-down approach to survey the field, focusing on fundamental challenges from three perspectives:

- Learning a world model to represent the visual environment and generalize to unseen ones.
- Learning a human model to effectively interpret human intentions from grounded instructions.
- Learning a VLN agent that leverages its world and human model to ground language, communicate, reason, and plan, enabling it to navigate environments as instructed.

We present a hierarchical and fine-grained taxonomy in Figure 2 to discuss challenges, solutions, and future directions based on foundation models for each model. To organize this survey, we start with a brief overview of the background and related research efforts as well as the available benchmarks in this field (§2). We structure the review around how the proposed methods have addressed the three key challenges described above: world model (§3), human model (§4), and VLN agent (§5). Finally, we discuss the current challenges and future research opportunities, particularly in light of the rise of foundation models (§6).

Name	World		Human			VLN Agent			Dataset	
	Domain	Environment	Turn	Format	Gran.	Type	Act. Sp.	Other	Text	Route
LANI/CHAI (2018)	Indoors	CHALET	Single	Multi Instr	A	-	Disc	Mani	H	H
R2R (2018)	Indoors	Matterport3D	Single	Multi Instr	A	Robot	Graph		H	P
R4R (2019)	Indoors	Matterport3D	Single	Multi Instr	A	Robot	Graph		H	P
RxR (2020)	Indoors	Matterport3D	Single	Multi Instr	A	Robot	Graph		H	P
SOON (2021a)	Indoors	Matterport3D	Single	Multi Instr	G	Robot	Graph		H	P
REVERIE (2020b)	Indoors	Matterport3D	Single	Multi Instr	A, G	Robot	Graph	Detect	H	P
VNLA (2019)	Indoors	Matterport3D	Multi	Multi Instr	A, G	Robot	Graph		T	P
HANNA (2019)	Indoors	Matterport3D	Multi	Multi Instr	A, G	Robot	Graph		H	P
CVDN (2020)	Indoors	Matterport3D	Multi	Restricted	A	Robot	Graph		H	H
VLN-CE (2020)	Indoors	Habitat, Matterport3D	Single	Multi Instr	A	Robot	Disc		H	P
Robo-VLN (2021)	Indoors	Habitat, Matterport3D	Single	Multi Instr	A	Robot	Cont		H	P
RobotSlang (2021)	Indoors	Real	Multi	Freeform	A	Robot	Disc		H	P
ALFRED (2020)	Indoors	AI2-THOR	Single	Multi Instr.	A, G	Robot	Disc	Mani	H	P
TEACh (2022)	Indoors	AI2-THOR	Multi	Freeform	A, G	Robot	Disc	Mani	H	H
DialFRED (2022)	Indoors	AI2-THOR	Multi	Restricted	A, G	Robot	Disc	Mani	H, T	P
TouchDown (2019)	Outdoors	Google Street View	Single	Multi Instr	A	-	Graph		H	P
Street Nav (2020)	Outdoors	Google Street View	Multi	Multi Instr	A	-	Disc		T	P
Talk2Nav (2021)	Outdoors	Google Street View	Single	Multi Instr	A, G	-	Disc		H	P
TtW (2018)	Outdoors	Real	Multi	Freeform	A, G	-	Disc		H	H
LCSD (2019)	Outdoors	CARLA	Single	Multi Instr	A	Driving	Disc		H	P
CDNLI (2020)	Outdoors	CARLA	Multi	Multi Instr	A, G	Driving	Cont		H, T	H
SDN (2022)	Outdoors	CARLA	Multi	Freeform	A, G	Driving	Disc, Cont		H	H
AerialVLN (2023b)	Outdoors	AirSim	Single	Multi Instr	A, G	Aerial	Disc		H	H
ANDH (2023a)	Outdoors	xView	Multi	Freeform	A, G	Aerial	Disc		H	H

Table 1: A summary of existing VLN benchmarks, taxonomized based on several key aspects: the **world** in which navigation occurs, the type of **human** interaction involved, the action space and tasks assigned to the **VLN agent**, and the methods of **dataset** collection. For the **world**, we consider their **domain** (either indoors or outdoors) and the **environment**. For the **human**, we consider their **turns of interaction** (either single or multiple turn), the **format of communication** (freeform dialogue, restricted dialogue, or multiple instructions), and the **language granularity** (action-directed and goal-directed). For the **VLN agent**, we consider their **agent types** (e.g., household robot, autonomous driving vehicles, or autonomous aerial vehicles), their **action space** (graph-based, discrete or continuous), and **other additional tasks** (manipulation and object detection). For **dataset collection**, we consider the **text collection** (by human or templated) and the **route demonstrations** (by human or planner).

2 Background and Task Formulations

In this section, we discuss the background, clarify the scope of this survey, define the VLN problem, and briefly overview the benchmarks.

2.1 Cognitive Underpinnings of VLN

Humans and other navigational animals demonstrate early understanding and strategies for navigating their environments (Rodrigo, 2002; Brand et al., 2015; Lingwood et al., 2018). For example, Gallistel (1990) describes two basic mechanisms: *piloting*, which involves environmental landmarks and computes distances and angles; and *path integration*, which calculates displacement and orientation changes through self-motion sensing. Central to understanding spatial navigation is the *cognitive map hypothesis*, suggesting that the brain forms a unified spatial representation to support memory and guide navigation (Epstein et al., 2017; Bellmund et al., 2018). For instance, Tolman (1948) observed that rats could adopt the correct novel path when familiar paths are blocked and landmarks are absent. Neuroscientists also discovered hippocampal place cells, indicating a spatial coordinate system that encodes landmarks and goals allocentrically (O'Keefe & Dostrovsky, 1971; O'keefe & Nadel, 1978). Recent studies also propose non-Euclidean representations, e.g., *cognitive graphs*, which illustrate the complexity of how we represent spatial knowledge of the world (Warren, 2019; Ericson & Warren, 2020). While visual and auditory perceptions are obviously integral to spatial representation (Klatzky et al., 2006), our linguistic skills and spatial cognition are also closely intertwined (Pruden et al., 2011). For instance, researchers have shown that understanding different aspects of spatial language can help with space-related tasks (Pyers et al., 2010), and that language influences how children interact with space by assisting them to recognize the importance of landmarks in identifying locations (Shusterman et al.,

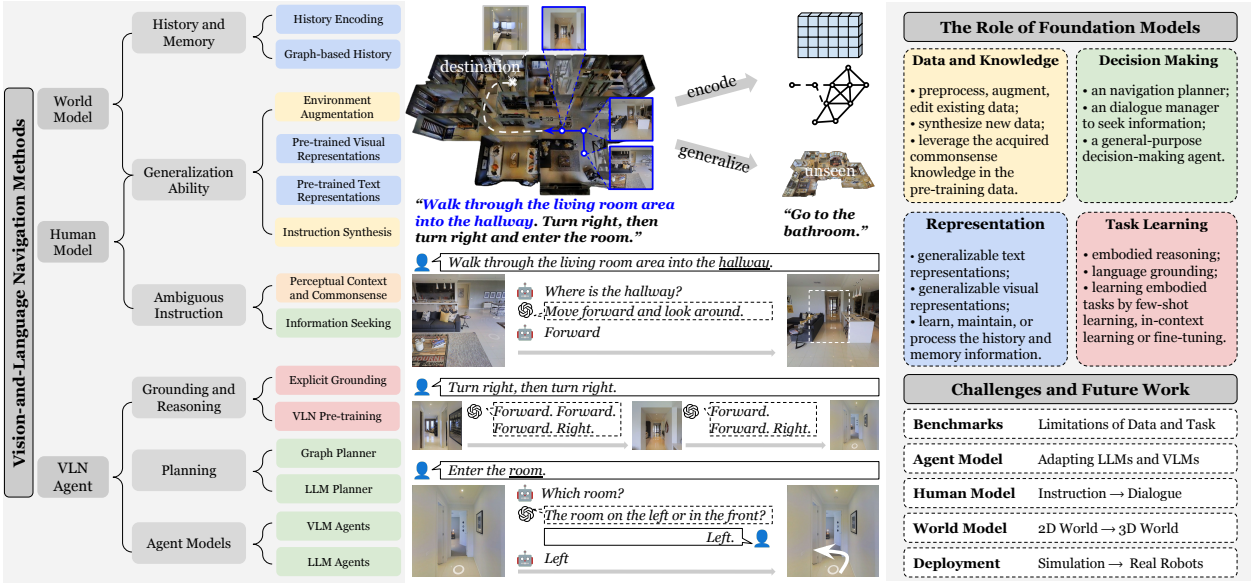

Figure 2: VLN challenges and solutions within the framework of world model, human model, and VLN agent. We discuss history and memory in the world model, ambiguous instructions in the human model, generalization ability in them both. For the VLN agent, we discuss methods for grounding and reasoning, planning, and adapting foundation models as agents. Depending on the role served by the foundation models, we categorize these methods into four types. Additionally, we discuss the potential future of the foundation model for the VLN task.

2011). Studying VLN not only enhances the development of embodied AI that follows human instructions in visual environments, but also deepens our understanding of how cognitive agents develop navigation skills, adapt to different environments, and how language use is connected to visual perceptions and actions.

2.2 Relevant Tasks and Scope of the Survey

Following natural language navigation instructions has traditionally been modeled using symbolic world representations such as maps (Anderson et al., 1991; MacMahon et al., 2006; Paz-Argaman & Tsarfaty, 2019). However, our survey focuses on models that employ visual environments and address the challenges of multimodal understanding and grounding. Likewise, we redirect readers to extensive surveys on visual navigation (Zhu et al., 2021b; Zhang et al., 2022a; Zhu et al., 2022) and mobile robot navigation (Gul et al., 2019; Crespo et al., 2020; Möller et al., 2021), which concentrate on visual perception and physical embodiment. However, these studies provide minimal discussions on the role of language in navigation tasks. While we inevitably extend our discussions of VLN to encompass areas beyond navigation, such as mobile manipulation and dialogue, our primary focus remains on navigational tasks, for which we provide a detailed literature review. Besides, unlike previous VLN surveys (Gu et al., 2022; Park & Kim, 2023; Wu et al., 2024), which offer a bottom-up summary focusing on benchmarks and modeling innovations, our survey adopts a top-down approach, and uses the roles of foundation models to categorize the research efforts into three fundamental challenges from the aspects of the world model, the human model, and the VLN agent. Note that this survey concentrates on frontier methods associated with the rise of foundation models. Thus, we point to less relevant methods (e.g., LSTM-based methods) very briefly at the beginning of each section to motivate our discussions.

2.3 VLN Task Formulations and Benchmarks

VLN Task Definition. A typical VLN agent receives a (sequence of) language instruction(s) from human instructors at a designated position. The agent navigates through the environment using an egocentric visual

perspective. By following the instructions, its task is to generate a trajectory over a sequence of discrete views or lower-level actions and control (e.g., FORWARD 0.25 meter) to reach the destination, which is considered successful if the agent arrives within a specified distance (e.g., 3 meters) from the destination. Besides, the agent may exchange information with the instructor during navigation, either by requesting help or engaging in freeform language communication. Additionally, there has been an increasing expectation for VLN agents to integrate additional tasks such as manipulation (Shridhar et al., 2020) and object detection (Qi et al., 2020b), along with navigation.

Benchmarks. Unlike other multimodal tasks such as VQA, which have a relatively fixed task definition and format, VLN encompasses a wide range of benchmarks and task formulations. These distinctions introduce unique challenges in addressing the broader VLN task and must be clearly understood as prerequisites for developing effective methods with appropriate foundation models. As is summarized in Table 1, existing VLN benchmarks can be taxonomized based on several key aspects in the LAW framework: (1) the **world** where navigation occurs, including the domain (indoors or outdoors) and the specifics of the environment. (2) the type of **human** interaction involved, including the interaction turns (single or multiple), communication format (freeform dialogue, restricted dialogue, or multiple instructions), and language granularity (action-directed or goal-directed). (3) the **VLN agent**, including its types (e.g., household robots, autonomous driving vehicles, or autonomous aerial vehicles), action space (graph-based, discrete, or continuous), and additional tasks (manipulation and object detection). (4) the **dataset** collection, including text collection method (human-generated or templated) and route demonstrations (human-performed or planner-generated). Representatively, Anderson et al. (2018) create the Room-to-Room (R2R) dataset based on the Matterport3D simulator (Chang et al., 2017), where an agent needs to follow fine-grained navigation instructions to reach the goal. Room-across-Room (RxR) (Ku et al., 2020) is a multilingual variation, including English, Hindi, and Telugu instructions. It offers a larger sample size and provides time-aligned instructions for virtual poses, enriching the task's linguistic and spatial information. Matterport3D allows VLN agents to operate in a *discrete environment* and rely on pre-defined connectivity graphs for navigation, where agents travel on the graph by teleportation between adjacent nodes, referred to as VLN-DE. To make the simplified setting more realistic, Krantz et al. (2020); Li et al. (2022c); Irshad et al. (2021) propose VLN in *continuous environments* (VLN-CE) by transferring discrete R2R paths to continuous spaces (Savva et al., 2019). Robo-VLN (Irshad et al., 2021) further narrows the sim-to-real gap by introducing VLN with continuous action spaces that are more realistic in robotics settings. Recent VLN benchmarks have undergone several design changes and expectations, which we discuss in § 6.

Evaluation Metrics. Three main metrics have been employed to evaluate navigation wayfinding performance (Anderson et al., 2018): (1) *Navigation Error (NE)*, the mean of the shortest path distance between the agent's final position and the goal destination; (2) *Success Rate (SR)*, the percentage of the final position being close enough to the goal destination; and (3) *Success Rate Weighted Path Length (SPL)*, which normalizes success rate by trajectory length to balance both the success rate in reaching the correct destination and the efficiency of the path. Some other metrics are used to measure the faithfulness of instruction following and the fidelity between the predicted and the ground-truth trajectory, for example: (4) *Coverage Weighted by Length Score (CLS)* (Jain et al., 2019)measures how closely an agent's trajectory follows the reference path. It balances two key aspects of the agent's performance: the extent of coverage of the reference path and the efficiency of the agent's navigation by considering the length score; (5) *Normalized Dynamic Time Warping (nDTW)* (Ilharco et al., 2019), which penalizes deviations from the ground-truth trajectories; and (6) *Normalized Dynamic Time Warping Weighted by Success Rate (sDTW)* (Ilharco et al., 2019), which penalizes deviations from the ground-truth trajectories and also considers the success rate.

2.4 Foundation Model

Foundation models are trained on large-scale datasets, which show strong generalization capability for a wide range of downstream applications. Text-only foundation models, such as pre-trained language models like BERT (Kenton & Toutanova, 2019) and GPT-3 (Brown, 2020), have revolutionized the field of NLP by setting new benchmarks for tasks like text generation, translation, and understanding. Building on the success of these models, vision-language (VL) foundation models, like LXMERT (Tan & Bansal, 2019),

CLIP (Radford et al., 2021) and GPT-4 (OpenAI, 2023), have expanded the paradigm to multimodal learning by integrating both visual and textual data, proving particularly impactful in various VL applications (Li et al., 2020a; 2019a; Ramesh et al., 2021; Alayrac et al., 2022; Hong et al., 2021). Some commonly used VL foundation models in VLN include LXMERT, which processes visual and textual inputs simultaneously, and CLIP (Radford et al., 2021), which uses a contrastive learning approach to align visual and textual data. For a more comprehensive overview of foundation models and their applications, we encourage readers to refer to existing survey papers such as Bommasani et al. (2021), Du et al. (2022), and Zhou et al. (2023).

3 World Model: Learning and Representing the Visual Environments

A world model helps the VLN agent to understand their surrounding environments, predict how their actions would change the world state, and align their perception and actions with language instructions. Two challenges have been highlighted in existing work about learning a world model: encoding the visual observation history in the current episode in the memory, and generalization to unseen environments in new tasks.

3.1 History and Memory

Different from other vision-language tasks like Visual Question Answering (VQA) (Antol et al., 2015), Visual Entailment (Xie et al., 2019), etc, the VLN agent needs to incorporate the history information of past actions and observations into its current step's input to determine the action rather than solely consider image and text in a single step. Prior to employing the foundation models in VLN, LSTM hidden states served as an implicit memory supporting agents' decision-making during navigation, and researchers further design different attention mechanisms (Tan et al., 2019; Wang et al., 2019) or auxiliary tasks (Ma et al., 2019; Zhu et al., 2020) to improve the alignment between the encoded history and instructions.

History Encoding. Different works have been proposed to encode navigation history using foundation models. A multi-modal Transformer is built upon encoded instructions and navigation history for decision-making, which is usually initialized from a model pre-trained on in-domain instruction-trajectory data like Prevalent (Hao et al., 2020). Some approaches encode the navigation history in recurrently updated state tokens. Hong et al. (2021) proposes to utilize a single `[CLS]` token from last step for encoding the history information, while Lin et al. (2022a) introduces a variable-length memory framework that stores multiple action activations from previous steps in a memory bank as the history encoding. Despite their effectiveness, these methods are limited by the need for step-by-step token updates, making it challenging to efficiently retrieve history encodings at arbitrary steps in the navigation trajectory, which can hinder scalability in pre-training. Another line of work directly encodes navigation history as a sequence with multi-modal Transformer. Among them, Pashevich et al. (2021) encodes single-view images for each step in a trajectory. Chen et al. (2021b) further proposes a panorama encoder to encode the panoramic visual observation at each time step, followed by a history encoder to encode all the past observations. This hierarchical design separately processes the spatial relationship in a panoramic view and the temporal dynamics across panoramas in the navigation history. Besides, this method eliminates the dependency on recurrently updated state tokens for history encoding, facilitating efficient and large-scale pre-training on instruction-path pairs. Follow-up research replaces the panorama encoder with mean pooling of images (Kamath et al., 2023) or front-view image encoding (Qiao et al., 2022), both maintaining effective navigation performance. With the advent of LLM-based navigation agents, some works (Zhou et al., 2024b) focus on converting the visual environment into textual descriptions, and explaining the world with text became the trend. The navigation history is then encoded as a sequence of these image descriptions, along with relative spatial information such as heading, elevation, and distance. HELPER (Sarch et al., 2023) equips an external memory of language-program pairs that parse free-form human-robot dialogue into action programs through retrieval-augmented LLM prompting.

Graph-based History. Another line of research enhances the navigation history modeling with graph information. For example, some works utilize a structured Transformer encoder to capture the geometric cues in the environment (Chen et al., 2022d; Deng et al., 2020; Wang et al., 2023b; Zhou & Mu, 2023; Su

et al., 2023; Zheng et al., 2023b; Wang et al., 2021; Chen et al., 2021a; Zhu et al., 2021a). In addition to the topological graph used in encoding, many works propose to include the top-down view information (e.g., grid map (Wang et al., 2023g; Liu et al., 2023a), semantic map (Hong et al., 2023a; Huang et al., 2023; Georgakis et al., 2022; Anderson et al., 2019; Chen et al., 2022b; Irshad et al., 2022), local metrics map (An et al., 2023)), and local neighborhood map (Gopinathan et al., 2023) in modeling the observation history during navigation. Recent advances in LLM-based navigation agents have introduced innovative approaches to memory construction using maps. For instance, Chen et al. (2024a) proposes a novel map-guided GPT-based agent that utilizes a linguistical-formed map to store and manage topological graph information. MC-GPT (Zhan et al., 2024b) introduces a topological map as the memory structure to record information about viewpoints, objects, and their spatial relationships.

3.2 Generalization across Environments

One main challenge in the VLN is learning from limited available environments and generalizing to new and unseen environments. Many works demonstrate that learning from semantic segmentation features (Zhang et al., 2020), dropout information in the environment during training (Tan et al., 2019), and maximizing the similarity between semantically-aligned image pairs from different environments (Li et al., 2022a) improve agents' generalization performance to unseen environments. These observations suggest the need to learn from large-scale environment data to avoid overfitting to training environments. Next, we discuss how existing works collect new environment data, and utilize it in training.

Pre-trained Visual Representations. Most works obtain vision representations from ResNet pre-trained on ImageNet (Anderson et al., 2018; Tan et al., 2019). Shen et al. (2021) replace it with the CLIP visual encoder (Radford et al., 2021), which is pre-trained with contrastive loss between image-text pairs and naturally aligns better with the instructions, boosting the VLN performance. Wang et al. (2022b) further explores transferring vision representation learned from video data for VLN task, suggesting that temporal information learned from video is crucial for navigation.

Environment Augmentation. One main line of research focuses on augmenting the navigation environment with auto-generated synthetic data. EnvEdit (Li et al., 2022b), EnvMix (Liu et al., 2021), KED (Zhu et al., 2023), and FDA (He et al., 2024a) generate synthetic data by changing the existing environments from Matterport3D. Specifically, they mix up rooms from different environments, change the appearance and style of the environments, and interpolate high-frequency features with the environments. Pathdreamer (Koh et al., 2021) and SE3DS (Koh et al., 2023) further synthesize the environments in future steps given current observations and explore utilizing the synthesis view as augmented data for VLN training.

The learning paradigm from the collected environments has changed with the advances in foundation models. Prior to the prevalence of pre-training in foundation models, most works directly augment the training environment with the auto-collected new environments and fine-tune a LSTM-based VLN agent (Li et al., 2022b; Liu et al., 2021; Koh et al., 2021; 2023; Zhu et al., 2023). As pre-training has been demonstrated to be crucial for foundation models, it has also become a standard practice in VLN to learn from collected environments during the pre-training stage (Li & Bansal, 2024; Kamath et al., 2023; Chen et al., 2022c; Wang et al., 2023h; Lin et al., 2023b; Guhur et al., 2021a; He et al., 2024a). Large-scale pre-training with augmented in-domain data has become crucial in bridging the gap between agents' and humans' performance. The in-domain pre-trained multi-modal transformer has been proven to be more effective than the multi-modal Transformer initialized from VLMs, like Oscar (Li et al., 2020b) and LXMERT (Tan & Bansal, 2019).

4 Human Model: Interpreting and Communicating with Humans

Besides learning and modeling the world, VLN agents also need a human model that comprehends human-provided natural language instructions per situation to complete navigation tasks. There are two main challenges: resolving ambiguity and generalization of grounded instructions in different visual environments.

4.1 Ambiguous Instructions

Ambiguous instructions mainly arise in single-turn navigation scenarios, where the agent follows an initial instruction without further human interaction for clarification. These instructions lack the flexibility to train the agent to adapt its language understanding and visual perception to the dynamic environment. For instance, instructions may contain landmarks invisible at the current view or indistinguishable landmarks visible from multiple views (Zhang & Kordjamshidi, 2023). The issue of ambiguous instructions is barely addressed before the application of foundational models to VLN. Although LEO (Xia et al., 2020) attempts to aggregate multiple instructions to describe the same trajectory from different perspectives, it still relies on human-annotated instructions. However, comprehensive perceptual context and commonsense knowledge from foundational models enable the agent to interpret ambiguous instructions using external knowledge, as well as seek assistance from other human models.

Perceptual Context and Commonsense Knowledge. Large-scale cross-modal pre-trained models like CLIP are capable of matching visual semantics with text. This enables the VLN agent to utilize information from the visual objects and their states in the current perception to resolve ambiguity, especially in single-turn navigation scenarios. For example, VLN-Trans (Zhang & Kordjamshidi, 2023) constructs easy-to-follow sub-instructions with visible and distinctive objects obtained from CLIP to pre-train a Translator that converts original ambiguous instructions into easily understandable sub-instruction representations. LANA+ (Wang et al., 2023f) leverages CLIP to query a text list of landmark semantic tags with the visual panoramic observations, and selects the top-ranked retrieved textual cues as representations of the salient landmarks to follow. NavHint (Zhang et al., 2024b) constructs a hint dataset, providing detailed visual descriptions to help the VLN agent build a comprehensive understanding of the visual environment rather than focusing solely on the objects mentioned in the instructions. On the other hand, the commonsense reasoning ability of LLMs can be used to clarify or correct ambiguous landmarks in the instructions, and break instructions into actionable items. For example, Lin et al. (2024b) use LLMs to provide commonsense about open-world landmark co-occurrences and conduct CLIP-driven landmark discovery accordingly. SayCan (Ahn et al., 2022) breaks an instruction into a ranked list of pre-defined admissible actions and combines them with an affordance function that assigns higher weights to the objects appearing in the current scene.

Information Seeking. While ambiguous instructions can be resolved based on visual perception and situational context, another more direct approach is to seek help from the communication partner, i.e., the human speakers who generate the instructions (Nguyen & Daumé III, 2019; Paul et al., 2022). There are three key challenges in this line of work: (1) deciding when to ask for help (Chi et al., 2020); (2) generating information-seeking questions, *e.g.*, next action, objects, and directions (Roman et al., 2020; Singh et al., 2022); (3) developing an oracle that provides the queried information, which could be either real humans (Singh et al., 2022), rules and templates (Gao et al., 2022), or neural models (Nguyen & Daumé III, 2019). LLMs and VLMs could potentially fit two roles in this framework, either as information-seeking models, or as proxies for human helpers or information-providing models. Preliminary research has explored the use of LLMs as the information-seeking model, addressing determining both when and what to ask. This is achieved with the help of techniques including conformal prediction (CP) (Ren et al., 2023) or in-context learning (ICL) (Chen et al., 2023c). For the latter, foundation models play the role of a helper who has access to oracle information, such as the location of the destination and a map of the environment, which is not available to the task performer. Very recently, VLN-Copilot (Qiao, 2024) enables agents to actively seek assistance when encountering confusion, with the LLM serving as a copilot to facilitate navigation. Fan et al. (2023b) demonstrate that GPT-3 can decompose ground-truth responses in the training data step-by-step, which helps in training an oracle model using a pre-trained SwinBert (Lin et al., 2022b) video-language model. They also demonstrate large vision-language models like mPLUG-Owl (Ye et al., 2023) can serve as strong zero-shot oracles off the shelf. Besides, self-motivated communication agents have been proposed (Zhu et al., 2021c) by learning the confidence of the oracle to produce a positive answer, which enables a self-Q&A manner where the oracle can be removed at inference time.

4.2 Generalization of Grounded Instructions

The limited scale and diversity of navigation data is another significant issue affecting the VLN agent's ability to comprehend various linguistic expressions and follow instructions effectively, particularly in unseen navigation environments. Although the language style itself has good generalization capability across seen and unseen environments (Zhang et al., 2020), how to ground the instructions with the unseen environments is potentially a hard task given the limited scale of training instructions. Foundation models help address these issues through both pre-trained representations and instruction generation for data augmentation.

Pre-trained Text Representations. Before the foundation models, many works rely on text encoders, such as LSTM, to represent text instructions (Anderson et al., 2018; Tan et al., 2019). The foundation models significantly enhance the VLN agent's language generalization ability through pre-trained representations. For example, PRESS (Li et al., 2019b) fine-tunes the pre-trained language model BERT (Kenton & Toutanova, 2019) to obtain text representations that generalize better to previously unseen instructions. The multi-modal Transformers (Tan & Bansal, 2019; Lu et al., 2019) boost methods, such as VLN-BERT (Majumdar et al., 2020) and PREVALENT (Hao et al., 2020), to obtain more generic vision-linguistic representations by pre-training on large-scale text-image pairs collected from the web. Airbert (Guhur et al., 2021b) trains ViLBERT-like architecture to learn text representations from image-caption pairs collected from the Internet. CLEAR (Li et al., 2022a) learns cross-lingual language representations that capture the visual concepts behind the instruction. ProbES (Liang et al., 2022) self-explores environments by sampling trajectories and automatically constructs the corresponding instruction by filling the instruction templates with movements and object phrases detected by CLIP. Additionally, it leverages prompt-based learning to facilitate fast adaptation of language embeddings. NavGPT-2 (Zhou et al., 2024b) explores leveraging vision-and-language representations from pre-trained VLMs (InstructBLIP (Dai et al., 2024) with Flan-T5 (Chung et al., 2024) or Vicuna (Zheng et al., 2023a)) to enhance policy learning for navigation and navigational reasoning.

Instruction Synthesis. Another method to improve the agent's generalization ability is to synthesize more instructions. Early works employ the Speaker-Follower framework (Fried et al., 2018; Tan et al., 2019; Kurita & Cho, 2020; Guhur et al., 2021a) to train an offline speaker (instruction generator) using human-annotated instruction-trajectory pairs. It then generates new instructions based on sequences of panoramas along a given trajectory. However, Zhao et al. (2021) observe that these generated instructions are low-quality and show a poor performance in human wayfinding evaluation. Marky (Wang et al., 2022a; Kamath et al., 2023) addresses this limitation using a multi-modal extension of the multilingual T5 model (Xue et al., 2020) with text-aligned visual landmark correspondences, achieving near-human quality on R2R-style paths in unseen environments. PASTS (Wang et al., 2023c) introduces a progress-aware spatial-temporal Transformer speaker to better leverage the sequenced multiple vision and action features. Additionally, instead of training an offline instruction generator, some recent research (Liang et al., 2022; Lin et al., 2023b; Zhang & Kordjamshidi, 2023; Wang et al., 2023e; Magassouba et al., 2021) generates instructions while navigating. For instance, LANA (Wang et al., 2023e) introduces a language-capable navigation agent that not only executes navigation instructions but also provides route descriptions.

5 VLN Agent: Learning an Embodied Agent for Reasoning and Planning

While the world and human models empower visual and language understanding abilities, VLN agents need to develop embodied reasoning and planning capabilities to support their decision-making. From this perspective, we discuss two challenges: grounding and reasoning, and planning. We also explore the method of directly applying foundation models as the VLN agent backbone.

5.1 Grounding and Reasoning

Different from other VL tasks, such as VQA and Image Captioning, which primarily focus on static alignment between images and corresponding textual descriptions, the VLN agent needs to reason about spatial and temporal dynamics in the instructions and the environment based on its actions. Specifically, the agent should consider previous actions, identify the part of the sub-instruction to execute, and ground the text

to the visual environment to execute the action accordingly. Previous methods primarily rely on explicit semantic modeling or auxiliary task design to obtain such abilities. However, pre-training with specially designed tasks has become the dominant approach with the advent of foundation models.

Explicit Semantic Grounding. The previous efforts enhance the agent's grounding ability through explicit semantic modeling in both vision and language modalities, including modeling motions and landmarks (Hong et al., 2020b;a; Zhang et al., 2021; Qi et al., 2020a), utilizing syntactic information in the instruction (Li et al., 2021), as well as spatial relations (Zhang & Kordjamshidi, 2022b). Very few works (Lin et al., 2023a; Zhan et al., 2024a; Wang et al., 2023b) explore explicit grounding in the VLN agent with the foundation models. Lin et al. (2023a) proposes actional atomic-concept learning and map visual observations to actional atomic concepts to facilitate multi-modal alignments.

Pre-training VLN Foundation Models. Except for explicit semantic modeling, the previous research also enhances the agent's grounding ability through auxiliary reasoning tasks (Ma et al., 2019; Wu et al., 2021; Zhu et al., 2020; Raychaudhuri et al., 2021; Dou & Peng, 2022; Kim et al., 2021). Such methods are less explored in VLN agents with foundation models, as their pre-training already provides a general understanding of spatial and temporal semantics prior to navigation. Various pre-training methods with specially designed tasks have been proposed to improve the agent's grounding ability. Lin et al. (2021) introduce pre-training tasks specifically designed for scene and object grounding. LOViS (Zhang & Kordjamshidi, 2022a) formulates two specialized pre-training tasks to enhance orientation and visual information separately. HOP (Qiao et al., 2022; 2023a) introduces a history-and-order aware pre-training paradigm that emphasizes historical information and trajectory orders. Li & Bansal (2023) suggests that enhancing the agent with the ability to predict future view semantics helps the agent in longer path navigation performance. Dou et al. (2023) design a masked path modeling objective to reconstruct the original path given a randomly masked sub-path. Cui et al. (2023) propose entity-aware pre-training by predicting grounded entities and aligning them to text.

5.2 Planning

Dynamic planning enables VLN agents to adapt to environmental changes and improve navigation strategies on the fly. Alongside the graph-based planners that utilize global graph information to enhance local action spaces, the rise of foundational models, particularly LLMs, has brought LLM-based planners into the VLN field. These planners use LLMs' vast commonsense knowledge and advanced reasoning to create dynamic plans that improve decision-making.

Graph-based Planner. Recent advancements in VLN emphasize enhancing navigational agents' planning capabilities through global graph information. Among them, Wang et al. (2021); Chen et al. (2022d); Deng et al. (2020); Zheng et al. (2023b) enhance the local navigation action spaces with global action steps from graph frontiers of visited nodes for better global planning. Gao et al. (2023) further enhances navigation decision-making with high-level planning for zone selection and low-level planning for node selection. Moreover, Liu et al. (2023a) enriches the graph-frontier-based global and local action spaces with grid-level actions for more accurate action prediction. In continuous environments, Krantz et al. (2021); Hong et al. (2022); Anderson et al. (2021) adopt a hierarchical planning approach utilizing high-level action spaces instead of low-level ones by selecting a local waypoint from a predicted local navigability graph. CM2 (Georgakis et al., 2022) facilitates trajectory planning by grounding instructions within a local map. Expanding on this strategy, An et al. (2024; 2023); Wang et al. (2023g); Chang et al. (2023b); Wang et al. (2022c) construct a global topological graph or grid maps to facilitate map-based global planning. Additionally, Wang et al. (2023a; 2024a) predict multiple future waypoints using either a video prediction model or a neural radiance representation model to plan the best action based on the long-term effects of predicted candidate waypoints.

LLM-based Planner. In parallel, some studies leverage common-sense knowledge from LLMs to generate text-based plans (Huang et al., 2022a;b). LLM-Planner (Song et al., 2022) creates detailed plans composed of sub-goals, dynamically adjusting these plans in real-time by integrating detected objects according to predefined program patterns. Similarly, Mic (Qiao et al., 2023b) and A^2Nav (Chen et al., 2023b) specialize

in breaking down navigation tasks into detailed textual instructions, with Mic generating step-by-step plans from both static and dynamic perspectives, while A^2Nav uses GPT-3 to parse instructions into actionable sub-tasks. ThinkBot (Lu et al., 2023) employs thought chain reasoning to generate missing actions with interactive objects. VL-Map (Huang et al., 2023) decomposes navigation instructions into sequential, goal-related functions in code format using code-written LLMs (following the Code-as-Policy (Liang et al., 2023) framework) and utilizes a dynamically built, queryable map to guide the execution of these goals. Additionally, SayNav (Rajvanshi et al., 2023) builds a 3D scene graph of the explored environment as input to LLMs for generating feasible and contextually appropriate high-level plans for the navigator.

5.3 Foundation Models as VLN Agents

The architecture of VLN agents has undergone significant transformations with the advent of foundation models. Initially conceptualized by Anderson et al. (2018), VLN agents were formulated within a Seq2Seq framework, employing an LSTM and an attention mechanism to model the interaction between vision and language modalities. With the advent of foundation models, the agent backend has transitioned from LSTM to Transformer and, more recently, to these large-scale pre-trained systems.

VLMs as Agents. The mainstream methodology leverages single-stream VLMs as the core structure of VLN agents (Hong et al., 2021; Qi et al., 2021; Moudgil et al., 2021; Zhao et al., 2022). These models process inputs from language, vision, and historical tokens simultaneously at each time step. It performs self-attention over these cross-modal tokens to capture the textual-visual correspondence, which is then used to infer the action probability. In the zero-shot VLN, CLIP-NAV (Dorbala et al., 2022) utilizes CLIP to obtain natural language referring expressions that describe the target object and make sequential navigational decisions. VLN-CE agents (Krantz et al., 2020) differentiate themselves from the VLN-DE (Anderson et al., 2018) agents by their action space, executing low-level controls in the continuous environment instead of graph-based high-level actions of view selection. Despite early works (Krantz et al., 2020; Raychaudhuri et al., 2021) utilizing LSTM to infer low-level actions, the introduction of waypoint predictors has allowed to transfer methods from DE to CE (Krantz et al., 2021; Krantz & Lee, 2022; Hong et al., 2022; Anderson et al., 2021). All these methods use a waypoint predictor to obtain a local navigability graph, allowing foundation models in DE to adapt to the continuous environment. In particular, the waypoint detection process primarily involves using visual observations (e.g., panoramic RGBD images) to predict navigable candidate adjacent waypoints from the agent's current position as possible targets. Given the predicted waypoints, the agent selects one as the current destination.

LLMs as Agents. Since LLMs have powerful reasoning ability and semantic abstraction of the world, and also show strong generalization ability in unknown large-scale environments, recent research in VLN has started to directly employ LLMs as agents to complete navigation. Typically, visual observations are converted into textual descriptions and fed into the LLM along with instructions, which then perform action predictions. Innovations such as NavGPT (Zhou et al., 2024a) and MapGPT (Chen et al., 2024a) demonstrate the feasibility of zero-shot navigation, with NavGPT autonomously generating actions using GPT-4 and MapGPT converting topological maps into global exploration hints. DiscussNav (Long et al., 2023) extends this approach by deploying multiple domain-specific VLN experts to automate and reduce human involvement in navigation tasks. It includes Instruction Analysis Experts, Vision Perception Experts, Completion Estimation Experts, and Decision Testing Experts. The use of multiple domain-specific VLN experts distributes tasks among specialized agents, reducing the burden on a single model and allowing for optimized, task-specific processing. This multi-expert approach enhances robustness, transparency, and overall performance by leveraging the collective strengths of multiple large models. MC-GPT (Zhan et al., 2024b) employs memory topology maps and human navigation examples to diversify strategies, while InstructNav (Long et al., 2024) breaks navigation into sub-tasks with multi-sourced value maps for effective execution. In contrast to zero-shot usage, some works (Zheng et al., 2024; Zhang et al., 2024a; Pan et al., 2023) fine-tune LLMs to address the embodied navigation tasks effectively. Some studies have incorporated the Chain-of-Thought (CoT) (Wei et al., 2022) reasoning mechanism to improve the reasoning process. Nav-CoT (Lin et al., 2024a) transforms LLMs into a world model and navigational reasoning agent, streamlining decisions by simulating future environments. This demonstrates the flexibility and practical potential of

fine-tuned language models in both simulation and real-world scenarios, marking a significant advancement over traditional applications.

6 Challenges and Future Directions

While foundation models have enabled novel solutions to VLN, several limitations remain underexplored and new challenges arise. In this section, we delve into the challenges and future direction of the VLN field from the perspectives of benchmarks, the world model, the human model, the agent model, and real robot deployment.

Benchmarks: Limitations of Data and Task. The current VLN datasets have limitations regarding quality, diversity, bias, and scalability. For example, in the R2R dataset, the instruction-trajectory pairs are biased to the shortest path, which may not accurately represent real-world navigation scenarios. We discuss the trends and recommendations on how VLN benchmarks can be improved.

- *Unified and Realistic Tasks and Platforms.* Establishing robust benchmarks and ensuring reproducibility are crucial for evaluating VLN in real-world settings. Real-world variability necessitates comprehensive benchmarks reflecting navigation challenges. A universal sim-to-real evaluation platform, like OVMM (Yenamandra et al., 2024), is needed for standardized testing across simulated and real-world settings. In addition, the tasks and activities should be realistic and designed originated from human needs. For instance, BEHAVIOR-1K (Li et al., 2024a) presents a benchmark of everyday household activities in virtual, interactive, and ecological environment to address the demands for diversity and realism.
- *Dynamic Environment.* Real-world environments are inherently complex and dynamic, with moving objects, people, and variations like lighting and weather presenting unexpected situations (Ma et al., 2022). These factors disrupt the visual perception of navigation systems and make maintaining reliable performance difficult. Recent efforts like HAZARD (Zhou et al., 2024c), Habitat 3.0 (Puig et al., 2024), and HA-VLN (Li et al., 2024c) consider dynamic environments and provide a good starting point.
- *Indoors to Outdoors.* VLN agents navigating in outdoor environments, e.g., autonomous driving and aerial vehicles, also start to get more attention (Vasudevan et al., 2021; Li et al., 2024b), with various language-guided datasets (Sriram et al., 2019; Ma et al., 2022) developed. Early studies have attempted to involve LLMs in these tasks, either with prompt engineering (Shah et al., 2023; Sha et al., 2023; Wen et al., 2023), or by fine-tuning LLMs to predict the next action or plan future trajectories (Chen et al., 2023a; Mao et al., 2023). To adapt off-the-shelf VLMs to these outdoor navigation domains, real-world driving videos (Xu et al., 2023; Yuan et al., 2024), simulated driving data (Wang et al., 2023d; Shao et al., 2023) and them both (Sima et al., 2023; Huang et al., 2024b) have been utilized for instruction tuning so that these foundation models learn to predict future throttle and steering angles. Additional reasoning and planning modules have also been integrated into foundation model driving agents (Huang et al., 2024b; Tian et al., 2024). We refer the readers to surveys and position papers for a detailed review (Li et al., 2023a; Cui et al., 2024; Gao et al., 2024; Yan et al., 2024).

World Model: From 2D to 3D. Building effective world representations is a central research theme in embodied perception, reasoning, and planning. Although the current research represents the world with strong and generic 2D representations, VLN is fundamentally a 3D task, where the agent perceives the real-world environment in 3D. Many explicit 3D representations are developed in prior work, including various semantic SLAMs and volumetric representation (Chaplot et al., 2020; Min et al., 2021; Saha et al., 2022; Blukis et al., 2022; Zhang et al., 2022b; Liu et al., 2024), depth information (An et al., 2023), Bird's-Eye-View representations like grid map (Wang et al., 2023g; Liu et al., 2023a), and local metrics map (An et al., 2023). These representations are limited because they reduce the object set to a closed set, making them inadequate for open-vocabulary settings with natural language. Several studies develop queryable map/scene representations by integrating multi-view image features captured from CLIP into 3D voxel grids (Jatavallabhula et al., 2023; Chang et al., 2023a) or top-down feature maps (Huang et al., 2023; Chen et al., 2022a), as well as utilizing scene graphs (Rana et al., 2023; Gu et al., 2023b) to represent spatial relationships. However, adapting 3D representations learned from large-scale data for VLN agents to better

perceive the 3D environment is still under exploration. The recent rise of 3D foundation models, including 3D reconstruction models (Hong et al., 2024) and 3D multimodal language models (Hong et al., 2023b; Yang et al., 2024; Huang et al., 2024a), can be crucial for VLN.

Human Model: From Instruction to Dialogue. Previous efforts predominantly adopt either a speaker-listener paradigm or restricted QA dialogue (Thomason et al., 2020; Gao et al., 2022) that only allows the agent to ask for help. Recently, there has been a surge in new benchmarks featuring open-ended dialogue instructions (De Vries et al., 2018; Banerjee et al., 2021; Padmakumar et al., 2022; Ma et al., 2022; Fan et al., 2023a), supporting fully free-form communication where agents can ask, propose, explain, suggest, clarify, and negotiate even in ambiguous or confusing scenarios. Still, current approaches rely on rule-based dialogue templates to tackle these complexities (Zhang et al., 2023; Parekh et al., 2023; Gu et al., 2023a), though they might feature a foundation model component. Huang et al. (2024b) perform conversational tuning on a video-language model using human-human dialogue data paired with simulated navigation videos, showcasing enhanced dialogue generation capabilities while navigation. Moving forward, it is imperative for future research to integrate foundation models for situated task-oriented dialogue management (Ulmer et al., 2024), or explore existing foundation models for task-oriented dialogue (He et al., 2022).

Agent Model: Adapting Foundation Models for VLN. While foundation models show strong generalizability, incorporating them into navigation tasks remains challenging. LLMs fundamentally lack the capability to visually perceive the actual environment and are prone to hallucinations. We als discuss capabilities of LLMs in planning and reasoning.

- *Lack of Embodied Experience.* This limitation can lead to scenarios where LLMs rely solely on pre-established commonsense for task planning and reasoning, which might not meet specific real-world needs (Xiang et al., 2024). Some pipelines tackle this issue by captioning the visual observations to textual descriptions as prompts for LLMs (Zheng et al., 2022), with a potential loss of essential visual semantics. Compared with LLMs, VLM agents demonstrate the potential to perceive the visual world and plan (Zhang et al., 2024a). Still, these models are primarily developed from internet data, which lack embodied experiences (Mu et al., 2024) and need finetuning for robust agentic decision-making (Zhai et al., 2024). Further research is needed to transfer the commonsense knowledge in foundation model agents to generalize to embodied situations. Recently proposed embodied foundation models (such as EmbodieGPT (Mu et al., 2024), PaLM-E (Driess et al., 2023) and Octopus (Yang et al., 2023)) offer a promising solution for enabling agents to operate more effectively in interactive environments. They fine-tune foundation models across multiple embodied tasks to bridge the gap between an agent's understanding of vision, language, and embodied actions, enhancing the foundation model's ability to comprehend and execute based on multimodal input.
- *Hallucination Issue.* LLMs and VLMs might generate non-existent objects, leading to misinformation (Li et al., 2023b; Chen et al., 2024b). For example, when LLM performs task planning, it may generate instructions such as "go forward and turn left at the sofa" even if there is no sofa in the room. This inaccuracy may cause them to execute incorrect or impossible actions.
- *LLMs in Planning and Reasoning.* There are some literatures evaluating the zero-shot reasoning and planning capabilities of LLMs, particularly in relation to PlanBench (Valmeekam et al., 2022) and CogEval (Momennejad et al., 2023), which highlight LLMs' limitations in more complex planning tasks. These works assess LLMs in a variety of challenging settings, such as plan generation, optimality, robustness, and reasoning, and identify that LLMs sometimes struggle with hallucinations or fail to grasp the relational structures underlying complex planning problems. In the context of VLN, the action space and the planning requirements are relatively constrained due to the fixed indoor environments and the limited set of navigational actions. This bounded setting makes it more feasible for LLMs to provide step-by-step instructions for coarse-grained directions, which has been demonstrated to be effective in previous works. In VLN tasks, the LLM's role is not to take over the entire planning process, but rather to assist by offering a structured breakdown of instructions. The agent's actual decision-making remains primarily reliant on other components, such as perception and motion control. Therefore, in VLN tasks, the LLM's planning serves more as a supplementary guide rather than the sole decision-making factor.

Deployment: From Simulation to Real Robots. Simulated settings often lack the complexity and variability of real-world environments, and lower-quality rendered images exacerbate this issue. First, the perception gap results in decreased performance and accuracy, highlighting the need for more robust perception systems. Wang et al. (2024b) have started to explore the use of semantic maps and 3D feature fields to provide monocular robots with panoramic perception shows improved performance. The embodiment gap and the data scarcity are also bottlenecks. The rise of robot teleportation (He et al., 2024b) also provides an alternative to scale up VLN data for foundation models in real human-robot communications.

7 Broader Impact

Foundation models hold great promise for advancing vision-language navigation. However, it is essential to address their broader ethical, legal, and societal implications. Given that they are pre-trained on vast, web-scale datasets, these models can carry inherent biases, which may result in fairness concerns, e.g., to multilingual users. Some approaches involve continual model training, it is critical to acknowledge and mitigate any potential risks to user privacy, especially when deployed in real-world applications such as home robotics.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *ArXiv preprint*, abs/2303.08774, 2023. URL `https://arxiv.org/abs/2303.08774`.

Michael Ahn, Anthony Brohan, Noah Brown, Yevgen Chebotar, Omar Cortes, Byron David, Chelsea Finn, Chuyuan Fu, Keerthana Gopalakrishnan, Karol Hausman, et al. Do as i can, not as i say: Grounding language in robotic affordances. *arXiv preprint arXiv:2204.01691*, 2022.

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35:23716–23736, 2022.

Dong An, Yuankai Qi, Yangguang Li, Yan Huang, Liang Wang, Tieniu Tan, and Jing Shao. Bevbert: Multimodal map pre-training for language-guided navigation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 2737–2748, 2023.

Dong An, Hanqing Wang, Wenguan Wang, Zun Wang, Yan Huang, Keji He, and Liang Wang. Etpnav: Evolving topological planning for vision-language navigation in continuous environments. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.

Anne H Anderson, Miles Bader, Ellen Gurman Bard, Elizabeth Boyle, Gwyneth Doherty, Simon Garrod, Stephen Isard, Jacqueline Kowtko, Jan McAllister, Jim Miller, et al. The hcrc map task corpus. *Language and speech*, 34(4):351–366, 1991.

Peter Anderson, Qi Wu, Damien Teney, Jake Bruce, Mark Johnson, Niko Sünderhauf, Ian Reid, Stephen Gould, and Anton Van Den Hengel. Vision-and-language navigation: Interpreting visually-grounded navigation instructions in real environments. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3674–3683, 2018.

Peter Anderson, Ayush Shrivastava, Devi Parikh, Dhruv Batra, and Stefan Lee. Chasing ghosts: Instruction following as bayesian state tracking. *Advances in neural information processing systems*, 32, 2019.

Peter Anderson, Ayush Shrivastava, Joanne Truong, Arjun Majumdar, Devi Parikh, Dhruv Batra, and Stefan Lee. Sim-to-real transfer for vision-and-language navigation. In *Conference on Robot Learning*, pp. 671–681. PMLR, 2021.

Jacob Andreas. Language models as agent models. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pp. 5769–5779, 2022.

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. VQA: Visual Question Answering. In *International Conference on Computer Vision (ICCV)*, 2015.

Shurjo Banerjee, Jesse Thomason, and Jason Corso. The robotslang benchmark: Dialog-guided robot localization and navigation. In *Conference on Robot Learning*, pp. 1384–1393. PMLR, 2021.

Jacob LS Bellmund, Peter Gärdenfors, Edvard I Moser, and Christian F Doeller. Navigating cognition: Spatial codes for human thinking. *Science*, 362(6415):eaat6766, 2018.

Valts Blukis, Chris Paxton, Dieter Fox, Animesh Garg, and Yoav Artzi. A persistent spatial semantic representation for high-level natural language instruction execution. In *Conference on Robot Learning*, pp. 706–717. PMLR, 2022.

Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*, 2021.

Rebecca J Brand, Kelly Escobar, Adrien Baranes, and Amanda Albu. Crawling predicts infants' understanding of agents' navigation of obstacles. *Infancy*, 20(4):405–415, 2015.

Tom B Brown. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*, 2020.

Angel Chang, Angela Dai, Thomas Funkhouser, Maciej Halber, Matthias Niessner, Manolis Savva, Shuran Song, Andy Zeng, and Yinda Zhang. Matterport3d: Learning from rgb-d data in indoor environments. *arXiv preprint arXiv:1709.06158*, 2017.

Haonan Chang, Kowndinya Boyalakuntla, Shiyang Lu, Siwei Cai, Eric Pu Jing, Shreesh Keskar, Shijie Geng, Adeeb Abbas, Lifeng Zhou, Kostas Bekris, et al. Context-aware entity grounding with open-vocabulary 3d scene graphs. In *Proceedings of the 2023 Conference on Robot Learning (CORL)*. JMLR, 2023a.

Matthew Chang, Theophile Gervet, Mukul Khanna, Sriram Yenamandra, Dhruv Shah, So Yeon Min, Kavit Shah, Chris Paxton, Saurabh Gupta, Dhruv Batra, et al. Goat: Go to any thing. *arXiv preprint arXiv:2311.06430*, 2023b.

Devendra Singh Chaplot, Dhiraj Prakashchand Gandhi, Abhinav Gupta, and Russ R Salakhutdinov. Object goal navigation using goal-oriented semantic exploration. *Advances in Neural Information Processing Systems*, 33:4247–4258, 2020.

Boyuan Chen, Fei Xia, Brian Ichter, Kanishka Rao, Keerthana Gopalakrishnan, Michael S. Ryoo, Austin Stone, and Daniel Kappler. Open-vocabulary queryable scene representations for real world planning. In *arXiv preprint arXiv:2209.09874*, 2022a.

Howard Chen, Alane Suhr, Dipendra Misra, Noah Snavely, and Yoav Artzi. Touchdown: Natural language navigation and spatial reasoning in visual street environments. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12538–12547, 2019.

Jiaqi Chen, Bingqian Lin, Ran Xu, Zhenhua Chai, Xiaodan Liang, and Kwan-Yee K Wong. Mapgpt: Map-guided prompting for unified vision-and-language navigation. *ArXiv preprint*, abs/2401.07314, 2024a. URL `https://arxiv.org/abs/2401.07314`.

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

Long Chen, Oleg Sinavski, Jan Hünermann, Alice Karnsund, Andrew James Willmott, Danny Birch, Daniel Maund, and Jamie Shotton. Driving with llms: Fusing object-level vector modality for explainable autonomous driving. *arXiv preprint arXiv:2310.01957*, 2023a.

Peihao Chen, Dongyu Ji, Kunyang Lin, Runhao Zeng, Thomas H. Li, Mingkui Tan, and Chuang Gan. Weakly-supervised multi-granularity map learning for vision-and-language navigation. In *Advances in Neural Information Processing Systems*, 2022b.

Peihao Chen, Xinyu Sun, Hongyan Zhi, Runhao Zeng, Thomas H. Li, Gaowen Liu, Mingkui Tan, and Chuang Gan. A^2nav: Action-aware zero-shot robot navigation by exploiting vision-and-language ability of foundation models. *CoRR*, abs/2308.07997, 2023b.

Shizhe Chen, Pierre-Louis Guhur, Cordelia Schmid, and Ivan Laptev. History aware multimodal transformer for vision-and-language navigation. In *Advances in Neural Information Processing Systems*, pp. 5834–5847, 2021b.

Shizhe Chen, Pierre-Louis Guhur, Makarand Tapaswi, Cordelia Schmid, and Ivan Laptev. Learning from unlabeled 3d environments for vision-and-language navigation. In *European Conference on Computer Vision*, pp. 638–655. Springer, 2022c.

Shizhe Chen, Pierre-Louis Guhur, Makarand Tapaswi, Cordelia Schmid, and Ivan Laptev. Think global, act local: Dual-scale graph transformer for vision-and-language navigation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16537–16547, 2022d.

Xiaoyu Chen, Shenao Zhang, Pushi Zhang, Li Zhao, and Jianyu Chen. Asking before acting: Gather information in embodied decision making with language models. *arXiv preprint arXiv:2305.15695*, 2023c.

Xuweiyi Chen, Ziqiao Ma, Xuejun Zhang, Sihan Xu, Shengyi Qian, Jianing Yang, David Fouhey, and Joyce Chai. Multi-object hallucination in vision language models. In *3rd Workshop on Advances in Language and Vision Research (ALVR)*, 2024b.

Ta-Chung Chi, Minmin Shen, Mihail Eric, Seokhwan Kim, and Dilek Hakkani-Tur. Just ask: An interactive learning framework for vision and language navigation. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pp. 2459–2466, 2020.

Xiangxiang Chu, Limeng Qiao, Xinyang Lin, Shuang Xu, Yang Yang, Yiming Hu, Fei Wei, Xinyu Zhang, Bo Zhang, Xiaolin Wei, and Chunhua Shen. Mobilevlm : A fast, strong and open vision language assistant for mobile devices. *ArXiv*, abs/2312.16886, 2023. URL `https://api.semanticscholar.org/CorpusID:266573855`.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. Scaling instruction-finetuned language models. *Journal of Machine Learning Research*, 25(70):1–53, 2024.

Jonathan Crespo, Jose Carlos Castillo, Oscar Martinez Mozos, and Ramon Barber. Semantic information for robot navigation: A survey. *Applied Sciences*, 10(2):497, 2020.

Can Cui, Yunsheng Ma, Xu Cao, Wenqian Ye, Yang Zhou, Kaizhao Liang, Jintai Chen, Juanwu Lu, Zichong Yang, Kuei-Da Liao, et al. A survey on multimodal large language models for autonomous driving. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 958–979, 2024.

Yibo Cui, Liang Xie, Yakun Zhang, Meishan Zhang, Ye Yan, and Erwei Yin. Grounded entity-landmark adaptive pre-training for vision-and-language navigation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 12043–12053, 2023.

Wenliang Dai, Junnan Li, DONGXU LI, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale N Fung, and Steven Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning. *Advances in Neural Information Processing Systems*, 36, 2024.

Harm De Vries, Kurt Shuster, Dhruv Batra, Devi Parikh, Jason Weston, and Douwe Kiela. Talk the walk: Navigating new york city through grounded dialogue. *arXiv preprint arXiv:1807.03367*, 2018.

Zhiwei Deng, Karthik Narasimhan, and Olga Russakovsky. Evolving graphical planner: Contextual global planning for vision-and-language navigation. *Advances in Neural Information Processing Systems*, 33: 20660–20672, 2020.

Vishnu Sashank Dorbala, Gunnar Sigurdsson, Robinson Piramuthu, Jesse Thomason, and Gaurav S Sukhatme. Clip-nav: Using clip for zero-shot vision-and-language navigation. *arXiv preprint arXiv:2211.16649*, 2022.

Zi-Yi Dou and Nanyun Peng. Foam: A follower-aware speaker model for vision-and-language navigation. *arXiv preprint arXiv:2206.04294*, 2022.

Zi-Yi Dou, Feng Gao, and Nanyun Peng. Masked path modeling for vision-and-language navigation. In *Findings of the Association for Computational Linguistics: EMNLP*, pp. 15255–15269, 2023.

Danny Driess, Fei Xia, Mehdi SM Sajjadi, Corey Lynch, Aakanksha Chowdhery, Brian Ichter, Ayzaan Wahid, Jonathan Tompson, Quan Vuong, Tianhe Yu, et al. Palm-e: An embodied multimodal language model. *arXiv preprint arXiv:2303.03378*, 2023.

Yifan Du, Zikang Liu, Junyi Li, and Wayne Xin Zhao. A survey of vision-language pre-trained models. *arXiv preprint arXiv:2202.10936*, 2022.

Jiafei Duan, Samson Yu, Hui Li Tan, Hongyuan Zhu, and Cheston Tan. A survey of embodied ai: From simulators to research tasks. *IEEE Transactions on Emerging Topics in Computational Intelligence*, 6(2): 230–244, 2022.

Russell A Epstein, Eva Zita Patai, Joshua B Julian, and Hugo J Spiers. The cognitive map in humans: spatial navigation and beyond. *Nature neuroscience*, 20(11):1504–1513, 2017.

Jonathan D Ericson and William H Warren. Probing the invariant structure of spatial knowledge: Support for the cognitive graph hypothesis. *Cognition*, 200:104276, 2020.

Yue Fan, Winson Chen, Tongzhou Jiang, Chun Zhou, Yi Zhang, and Xin Wang. Aerial vision-and-dialog navigation. In *Findings of the Association for Computational Linguistics: ACL 2023*, pp. 3043–3061, 2023a.

Yue Fan, Jing Gu, Kaizhi Zheng, and Xin Wang. R2h: Building multimodal navigation helpers that respond to help requests. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 14803–14819, 2023b.

Daniel Fried, Ronghang Hu, Volkan Cirik, Anna Rohrbach, Jacob Andreas, Louis-Philippe Morency, Taylor Berg-Kirkpatrick, Kate Saenko, Dan Klein, and Trevor Darrell. Speaker-follower models for vision-and-language navigation. *Advances in neural information processing systems*, 31, 2018.

Charles R Gallistel. *The organization of learning.* The MIT Press, 1990.

Chen Gao, Xingyu Peng, Mi Yan, He Wang, Lirong Yang, Haibing Ren, Hongsheng Li, and Si Liu. Adaptive zone-aware hierarchical planner for vision-language navigation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14911–14920, 2023.

Haoxiang Gao, Yaqian Li, Kaiwen Long, Ming Yang, and Yiqing Shen. A survey for foundation models in autonomous driving. *arXiv preprint arXiv:2402.01105*, 2024.

Xiaofeng Gao, Qiaozi Gao, Ran Gong, Kaixiang Lin, Govind Thattai, and Gaurav S Sukhatme. Dialfred: Dialogue-enabled agents for embodied instruction following. *IEEE Robotics and Automation Letters (RA-L)*, 2022.

Georgios Georgakis, Karl Schmeckpeper, Karan Wanchoo, Soham Dan, Eleni Miltsakaki, Dan Roth, and Kostas Daniilidis. Cross-modal map learning for vision and language navigation. In *Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition*, pp. 15439–15449, 2022.

Muraleekrishna Gopinathan, Jumana Abu-Khalaf, David Suter, Sidike Paheding, and Nathir A Rawashdeh. What is near?: Room locality learning for enhanced robot vision-language-navigation in indoor living environments. *arXiv preprint arXiv:2309.05036*, 2023.

Jing Gu, Eliana Stefani, Qi Wu, Jesse Thomason, and Xin Eric Wang. Vision-and-language navigation: A survey of tasks, methods, and future directions. *arXiv preprint arXiv:2203.12667*, 2022.

Jing Gu, Kaizhi Zheng, Kaiwen Zhou Yue Fan, Xuehai He Jialu Wang Zonglin Di, and Xin Eric Wang. Slugjarvis: Multimodal commonsense knowledge-based embodied ai for simbot challenge. In *Alexa Prize SimBot Challenge Proceedings*, 2023a.

Qiao Gu, Alihusein Kuwajerwala, Sacha Morin, Krishna Murthy Jatavallabhula, Bipasha Sen, Aditya Agarwal, Corban Rivera, William Paul, Kirsty Ellis, Rama Chellappa, et al. Conceptgraphs: Open-vocabulary 3d scene graphs for perception and planning. *arXiv preprint arXiv:2309.16650*, 2023b.

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

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

Faiza Gul, Wan Rahiman, and Syed Sahal Nazli Alhady. A comprehensive study for robot navigation techniques. *Cogent Engineering*, 6(1):1632046, 2019.

David Ha and Jürgen Schmidhuber. Recurrent world models facilitate policy evolution. In *Advances in Neural Information Processing Systems 31*, pp. 2451–2463. 2018.

Weituo Hao, Chunyuan Li, Xiujun Li, Lawrence Carin, and Jianfeng Gao. Towards learning a generic agent for vision-and-language navigation via pre-training. In *Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition*, pp. 13134–13143, 2020.

Keji He, Chenyang Si, Zhihe Lu, Yan Huang, Liang Wang, and Xinchao Wang. Frequency-enhanced data augmentation for vision-and-language navigation. *Advances in Neural Information Processing Systems*, 36, 2024a.

Tairan He, Zhengyi Luo, Wenli Xiao, Chong Zhang, Kris Kitani, Changliu Liu, and Guanya Shi. Learning human-to-humanoid real-time whole-body teleoperation. *arXiv preprint arXiv:2403.04436*, 2024b.

Wanwei He, Yinpei Dai, Yinhe Zheng, Yuchuan Wu, Zheng Cao, Dermot Liu, Peng Jiang, Min Yang, Fei Huang, Luo Si, et al. Galaxy: A generative pre-trained model for task-oriented dialog with semi-supervised learning and explicit policy injection. In *Proceedings of the AAAI conference on artificial intelligence*, volume 36, pp. 10749–10757, 2022.

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

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

Yicong Hong, Cristian Rodriguez-Opazo, Qi Wu, and Stephen Gould. Sub-instruction aware vision-and-language navigation. *arXiv preprint arXiv:2004.02707*, 2020b.

Yicong Hong, Qi Wu, Yuankai Qi, Cristian Rodriguez Opazo, and Stephen Gould. A recurrent vision-and-language BERT for navigation. In *Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition*, 2021.

Yicong Hong, Zun Wang, Qi Wu, and Stephen Gould. Bridging the gap between learning in discrete and continuous environments for vision-and-language navigation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15439–15449, 2022.

Yicong Hong, Yang Zhou, Ruiyi Zhang, Franck Dernoncourt, Trung Bui, Stephen Gould, and Hao Tan. Learning navigational visual representations with semantic map supervision. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 3055–3067, 2023a.

Yicong Hong, Kai Zhang, Jiuxiang Gu, Sai Bi, Yang Zhou, Difan Liu, Feng Liu, Kalyan Sunkavalli, Trung Bui, and Hao Tan. LRM: Large reconstruction model for single image to 3d. In *The Twelfth International Conference on Learning Representations*, 2024.

Yining Hong, Haoyu Zhen, Peihao Chen, Shuhong Zheng, Yilun Du, Zhenfang Chen, and Chuang Gan. 3d-llm: Injecting the 3d world into large language models. *NeurIPS*, 2023b.

Yihan Hu, Jiazhi Yang, Li Chen, Keyu Li, Chonghao Sima, Xizhou Zhu, Siqi Chai, Senyao Du, Tianwei Lin, Wenhai Wang, Lewei Lu, Xiaosong Jia, Qiang Liu, Jifeng Dai, Yu Qiao, and Hongyang Li. Planning-oriented autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023.

Zhiting Hu and Tianmin Shu. Language models, agent models, and world models: The law for machine reasoning and planning. *arXiv preprint arXiv:2312.05230*, 2023.

Chenguang Huang, Oier Mees, Andy Zeng, and Wolfram Burgard. Visual language maps for robot navigation. In *IEEE International Conference on Robotics and Automation*, pp. 10608–10615. IEEE, 2023.

Jiangyong Huang, Silong Yong, Xiaojian Ma, Xiongkun Linghu, Puhao Li, Yan Wang, Qing Li, Song-Chun Zhu, Baoxiong Jia, and Siyuan Huang. An embodied generalist agent in 3d world. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2024a.

Wenlong Huang, Pieter Abbeel, Deepak Pathak, and Igor Mordatch. Language models as zero-shot planners: Extracting actionable knowledge for embodied agents. In *International Conference on Machine Learning*, pp. 9118–9147. PMLR, 2022a.

Wenlong Huang, Fei Xia, Ted Xiao, Harris Chan, Jacky Liang, Pete Florence, Andy Zeng, Jonathan Tompson, Igor Mordatch, Yevgen Chebotar, Pierre Sermanet, Noah Brown, Tomas Jackson, Linda Luu, Sergey Levine, Karol Hausman, and Brian Ichter. Inner monologue: Embodied reasoning through planning with language models. In *arXiv preprint arXiv:2207.05608*, 2022b.

Yidong Huang, Jacob Sansom, Ziqiao Ma, Felix Gervits, and Joyce Chai. Drivlme: Enhancing llm-based autonomous driving agents with embodied and social experiences. In *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2024b.

Gabriel Ilharco, Vihan Jain, Alexander Ku, Eugene Ie, and Jason Baldridge. General evaluation for instruction conditioned navigation using dynamic time warping. In *Advances in neural information processing systems*, 2019.

Muhammad Zubair Irshad, Chih-Yao Ma, and Zsolt Kira. Hierarchical cross-modal agent for robotics vision-and-language navigation. In *2021 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 13238–13246. IEEE, 2021.

Muhammad Zubair Irshad, Niluthpol Chowdhury Mithun, Zachary Seymour, Han-Pang Chiu, Supun Samarasekera, and Rakesh Kumar. Semantically-aware spatio-temporal reasoning agent for vision-and-language navigation in continuous environments. In *International Conference on Pattern Recognition*, pp. 4065–4071. IEEE, 2022.

Vihan Jain, Gabriel Magalhaes, Alexander Ku, Ashish Vaswani, Eugene Ie, and Jason Baldridge. Stay on the path: Instruction fidelity in vision-and-language navigation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 1862–1872, 2019.

Krishna Murthy Jatavallabhula, Alihusein Kuwajerwala, Qiao Gu, Mohd Omama, Tao Chen, Alaa Maalouf, Shuang Li, Ganesh Iyer, Soroush Saryazdi, Nikhil Keetha, et al. Conceptfusion: Open-set multimodal 3d mapping. *arXiv preprint arXiv:2302.07241*, 2023.

Aishwarya Kamath, Peter Anderson, Su Wang, Jing Yu Koh, Alexander Ku, Austin Waters, Yinfei Yang, Jason Baldridge, and Zarana Parekh. A new path: Scaling vision-and-language navigation with synthetic instructions and imitation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10813–10823, 2023.

Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. pp. 4171–4186, 2019.

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

Roberta L Klatzky, James R Marston, Nicholas A Giudice, Reginald G Golledge, and Jack M Loomis. Cognitive load of navigating without vision when guided by virtual sound versus spatial language. *Journal of experimental psychology: Applied*, 12(4):223, 2006.

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

Jing Yu Koh, Harsh Agrawal, Dhruv Batra, Richard Tucker, Austin Waters, Honglak Lee, Yinfei Yang, Jason Baldridge, and Peter Anderson. Simple and effective synthesis of indoor 3d scenes. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 1169–1178, 2023.

Jacob Krantz and Stefan Lee. Sim-2-sim transfer for vision-and-language navigation in continuous environments. In *European Conference on Computer Vision*, pp. 588–603. Springer, 2022.

Jacob Krantz, Erik Wijmans, Arjun Majumdar, Dhruv Batra, and Stefan Lee. Beyond the nav-graph: Vision-and-language navigation in continuous environments. In *European Conference on Computer Vision*, pp. 104–120. Springer, 2020.

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

Alexander Ku, Peter Anderson, Roma Patel, Eugene Ie, and Jason Baldridge. Room-across-room: Multi-lingual vision-and-language navigation with dense spatiotemporal grounding. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 4392–4412, 2020.

Shuhei Kurita and Kyunghyun Cho. Generative language-grounded policy in vision-and-language navigation with bayes' rule. *arXiv preprint arXiv:2009.07783*, 2020.

Chengshu Li, Ruohan Zhang, Josiah Wong, Cem Gokmen, Sanjana Srivastava, Roberto Martín-Martín, Chen Wang, Gabrael Levine, Wensi Ai, Benjamin Martinez, et al. Behavior-1k: A human-centered, embodied ai benchmark with 1,000 everyday activities and realistic simulation. *arXiv preprint arXiv:2403.09227*, 2024a.

Gen Li, Nan Duan, Yuejian Fang, Ming Gong, and Daxin Jiang. Unicoder-vl: A universal encoder for vision and language by cross-modal pre-training. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pp. 11336–11344, 2020a.

Jialu Li and Mohit Bansal. Improving vision-and-language navigation by generating future-view image semantics. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10803–10812, 2023.

Jialu Li and Mohit Bansal. Panogen: Text-conditioned panoramic environment generation for vision-and-language navigation. *Advances in Neural Information Processing Systems*, 36, 2024.

Jialu Li, Hao Tan, and Mohit Bansal. Improving cross-modal alignment in vision language navigation via syntactic information. *arXiv preprint arXiv:2104.09580*, 2021.

Jialu Li, Hao Tan, and Mohit Bansal. Clear: Improving vision-language navigation with cross-lingual, environment-agnostic representations. *arXiv preprint arXiv:2207.02185*, 2022a.

Jialu Li, Hao Tan, and Mohit Bansal. Envedit: Environment editing for vision-and-language navigation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15407–15417, 2022b.

Jialu Li, Aishwarya Padmakumar, Gaurav Sukhatme, and Mohit Bansal. Vln-video: Utilizing driving videos for outdoor vision-and-language navigation. *arXiv preprint arXiv:2402.03561*, 2024b.

Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. Visualbert: A simple and performant baseline for vision and language. *arXiv preprint arXiv:1908.03557*, 2019a.

Minghan Li, Heng Li, Zhi-Qi Cheng, Yifei Dong, Yuxuan Zhou, Jun-Yan He, Qi Dai, Teruko Mitamura, and Alexander G. Hauptmann. Human-aware vision-and-language navigation: Bridging simulation to reality with dynamic human interactions, 2024c. URL `https://arxiv.org/abs/2406.19236`.

Xin Li, Yeqi Bai, Pinlong Cai, Licheng Wen, Daocheng Fu, Bo Zhang, Xuemeng Yang, Xinyu Cai, Tao Ma, Jianfei Guo, et al. Towards knowledge-driven autonomous driving. *arXiv preprint arXiv:2312.04316*, 2023a.

Xinghang Li, Di Guo, Huaping Liu, and Fuchun Sun. Reve-ce: Remote embodied visual referring expression in continuous environment. *IEEE Robotics and Automation Letters*, 7(2):1494–1501, 2022c.

Xiujun Li, Chunyuan Li, Qiaolin Xia, Yonatan Bisk, Asli Çelikyilmaz, Jianfeng Gao, Noah A. Smith, and Yejin Choi. Robust navigation with language pretraining and stochastic sampling. 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 2019, Hong Kong, China, November 3-7, 2019*, pp. 1494–1499, 2019b.

Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, et al. Oscar: Object-semantics aligned pre-training for vision-language tasks. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXX 16*, pp. 121–137. Springer, 2020b.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 2023b.

Jacky Liang, Wenlong Huang, Fei Xia, Peng Xu, Karol Hausman, Brian Ichter, Pete Florence, and Andy Zeng. Code as policies: Language model programs for embodied control. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 9493–9500. IEEE, 2023.

Xiwen Liang, Fengda Zhu, Lingling Li, Hang Xu, and Xiaodan Liang. Visual-language navigation pretraining via prompt-based environmental self-exploration. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 4837–4851, 2022.

Bingqian Lin, Yi Zhu, Xiaodan Liang, Liang Lin, and Jianzhuang Liu. Actional atomic-concept learning for demystifying vision-language navigation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 1568–1576, 2023a.

Bingqian Lin, Yunshuang Nie, Ziming Wei, Jiaqi Chen, Shikui Ma, Jianhua Han, Hang Xu, Xiaojun Chang, and Xiaodan Liang. Navcot: Boosting llm-based vision-and-language navigation via learning disentangled reasoning. *arXiv preprint arXiv:2403.07376*, 2024a.

Bingqian Lin, Yunshuang Nie, Ziming Wei, Yi Zhu, Hang Xu, Shikui Ma, Jianzhuang Liu, and Xiaodan Liang. Correctable landmark discovery via large models for vision-language navigation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024b.

Chuang Lin, Yi Jiang, Jianfei Cai, Lizhen Qu, Gholamreza Haffari, and Zehuan Yuan. Multimodal transformer with variable-length memory for vision-and-language navigation. In *European Conference on Computer Vision*, volume 13696, pp. 380–397. Springer, 2022a.

Kevin Lin, Linjie Li, Chung-Ching Lin, Faisal Ahmed, Zhe Gan, Zicheng Liu, Yumao Lu, and Lijuan Wang. Swinbert: End-to-end transformers with sparse attention for video captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 17949–17958, 2022b.

Kunyang Lin, Peihao Chen, Diwei Huang, Thomas H Li, Mingkui Tan, and Chuang Gan. Learning vision-and-language navigation from youtube videos. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 8317–8326, 2023b.

Xiangru Lin, Guanbin Li, and Yizhou Yu. Scene-intuitive agent for remote embodied visual grounding. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 7036–7045, 2021.

Jamie Lingwood, Mark Blades, Emily K Farran, Yannick Courbois, and Danielle Matthews. Using virtual environments to investigate wayfinding in 8-to 12-year-olds and adults. *Journal of experimental child psychology*, 166:178–189, 2018.

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

Rui Liu, Xiaohan Wang, Wenguan Wang, and Yi Yang. Bird's-eye-view scene graph for vision-language navigation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 10968–10980, 2023a.

Rui Liu, Wenguan Wang, and Yi Yang. Volumetric environment representation for vision-language navigation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16317–16328, 2024.

Shubo Liu, Hongsheng Zhang, Yuankai Qi, Peng Wang, Yanning Zhang, and Qi Wu. Aerialvln: Vision-and-language navigation for uavs. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 15384–15394, 2023b.

Yuxing Long, Xiaoqi Li, Wenzhe Cai, and Hao Dong. Discuss before moving: Visual language navigation via multi-expert discussions. *ArXiv preprint*, abs/2309.11382, 2023. URL `https://arxiv.org/abs/2309.11382`.

Yuxing Long, Wenzhe Cai, Hongcheng Wang, Guanqi Zhan, and Hao Dong. Instructnav: Zero-shot system for generic instruction navigation in unexplored environment, 2024. URL `https://arxiv.org/abs/2406.04882`.

Guanxing Lu, Ziwei Wang, Changliu Liu, Jiwen Lu, and Yansong Tang. Thinkbot: Embodied instruction following with thought chain reasoning. *arXiv preprint arXiv:2312.07062*, 2023.

Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *Advances in neural information processing systems*, 32, 2019.

Chih-Yao Ma, Jiasen Lu, Zuxuan Wu, Ghassan AlRegib, Zsolt Kira, Richard Socher, and Caiming Xiong. Self-monitoring navigation agent via auxiliary progress estimation. *arXiv preprint arXiv:1901.03035*, 2019.

Ziqiao Ma, Benjamin VanDerPloeg, Cristian-Paul Bara, Yidong Huang, Eui-In Kim, Felix Gervits, Matthew Marge, and Joyce Chai. Dorothie: Spoken dialogue for handling unexpected situations in interactive autonomous driving agents. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pp. 4800–4822, 2022.

Ziqiao Ma, Jacob Sansom, Run Peng, and Joyce Chai. Towards a holistic landscape of situated theory of mind in large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 1011–1031, 2023.

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

Aly Magassouba, Komei Sugiura, and Hisashi Kawai. Crossmap transformer: A crossmodal masked path transformer using double back-translation for vision-and-language navigation. *IEEE Robotics and Automation Letters*, 6:6258–6265, 2021. URL `https://api.semanticscholar.org/CorpusID:232075933`.

Arjun Majumdar, Ayush Shrivastava, Stefan Lee, Peter Anderson, Devi Parikh, and Dhruv Batra. Improving vision-and-language navigation with image-text pairs from the web. In *European Conference on Computer Vision*, pp. 259–274, 2020.

Jiageng Mao, Yuxi Qian, Junjie Ye, Hang Zhao, and Yue Wang. Gpt-driver: Learning to drive with gpt. In *NeurIPS 2023 Foundation Models for Decision Making Workshop*, 2023.

So Yeon Min, Devendra Singh Chaplot, Pradeep Kumar Ravikumar, Yonatan Bisk, and Ruslan Salakhutdinov. Film: Following instructions in language with modular methods. In *International Conference on Learning Representations*, 2021.

Piotr Mirowski, Matt Grimes, Mateusz Malinowski, Karl Moritz Hermann, Keith Anderson, Denis Teplyashin, Karen Simonyan, Andrew Zisserman, Raia Hadsell, et al. Learning to navigate in cities without a map. *Advances in neural information processing systems*, 31, 2018.

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

Ronja Möller, Antonino Furnari, Sebastiano Battiato, Aki Härmä, and Giovanni Maria Farinella. A survey on human-aware robot navigation. *Robotics and Autonomous Systems*, 145:103837, 2021.

Ida Momennejad, Hosein Hasanbeig, Felipe Vieira Frujeri, Hiteshi Sharma, Robert Osazuwa Ness, Nebojsa Jojic, Hamid Palangi, and Jonathan Larson. Evaluating cognitive maps and planning in large language models with cogeval. *ArXiv*, abs/2309.15129, 2023.

Abhinav Moudgil, Arjun Majumdar, Harsh Agrawal, Stefan Lee, and Dhruv Batra. SOAT: A scene- and object-aware transformer for vision-and-language navigation. In *Advances in neural information processing systems*, pp. 7357–7367, 2021.

Yao Mu, Qinglong Zhang, Mengkang Hu, Wenhai Wang, Mingyu Ding, Jun Jin, Bin Wang, Jifeng Dai, Yu Qiao, and Ping Luo. Embodiedgpt: Vision-language pre-training via embodied chain of thought. *Advances in Neural Information Processing Systems*, 36, 2024.

Khanh Nguyen and Hal Daumé III. 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)*, pp. 684–695, Hong Kong, China, November 2019.

Khanh Nguyen, Debadeepta Dey, Chris Brockett, and Bill Dolan. Vision-based navigation with language-based assistance via imitation learning with indirect intervention. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12527–12537, 2019.

Phuong DH Nguyen, Yasmin Kim Georgie, Ezgi Kayhan, Manfred Eppe, Verena Vanessa Hafner, and Stefan Wermter. Sensorimotor representation learning for an "active self" in robots: a model survey. *KI-Künstliche Intelligenz*, 35:9–35, 2021.

John O'Keefe and Jonathan Dostrovsky. The hippocampus as a spatial map: preliminary evidence from unit activity in the freely-moving rat. *Brain research*, 1971.

John O'keefe and Lynn Nadel. *The hippocampus as a cognitive map*. Oxford university press, 1978.

OpenAI. Gpt-4 technical report, 2023. URL `https://cdn.openai.com/papers/gpt-4.pdf`.

Aishwarya Padmakumar, Jesse Thomason, Ayush Shrivastava, Patrick Lange, Anjali Narayan-Chen, Spandana Gella, Robinson Piramuthu, Gokhan Tur, and Dilek Hakkani-Tur. Teach: Task-driven embodied agents that chat. In *AAAI*, 2022.

Bowen Pan, Rameswar Panda, SouYoung Jin, Rogerio Feris, Aude Oliva, Phillip Isola, and Yoon Kim. Langnav: Language as a perceptual representation for navigation. *arXiv preprint arXiv:2310.07889*, 2023.

Amit Parekh, Malvina Nikandrou, Georgios Pantazopoulos, Bhathiya Hemanthage, Arash Eshghi, Ioannis Konstas, Oliver Lemon, and Alessandro Suglia. Emma: A foundation model for embodied, interactive, multimodal task completion in 3d environments. In *Alexa Prize SimBot Challenge Proceedings*, 2023.

Sang-Min Park and Young-Gab Kim. Visual language navigation: A survey and open challenges. *Artificial Intelligence Review*, 56(1):365–427, 2023.

Alexander Pashevich, Cordelia Schmid, and Chen Sun. Episodic transformer for vision-and-language navigation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 15922–15932. IEEE, 2021.

Sudipta Paul, Amit Roy-Chowdhury, and Anoop Cherian. Avlen: Audio-visual-language embodied navigation in 3d environments. *Advances in Neural Information Processing Systems*, 35:6236–6249, 2022.

Tzuf Paz-Argaman and Reut Tsarfaty. RUN through the streets: A new dataset and baseline models for realistic urban navigation. 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)*, pp. 6449–6455, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1681. URL `https://aclanthology.org/D19-1681`.

Shannon M Pruden, Susan C Levine, and Janellen Huttenlocher. Children's spatial thinking: Does talk about the spatial world matter? *Developmental science*, 14(6):1417–1430, 2011.

Xavier Puig, Eric Undersander, Andrew Szot, Mikael Dallaire Cote, Tsung-Yen Yang, Ruslan Partsey, Ruta Desai, Alexander Clegg, Michal Hlavac, So Yeon Min, et al. Habitat 3.0: A co-habitat for humans, avatars, and robots. In *The Twelfth International Conference on Learning Representations*, 2024.

Jennie E Pyers, Anna Shusterman, Ann Senghas, Elizabeth S Spelke, and Karen Emmorey. Evidence from an emerging sign language reveals that language supports spatial cognition. *Proceedings of the National Academy of Sciences*, 107(27):12116–12120, 2010.

Yuankai Qi, Zizheng Pan, Shengping Zhang, Anton van den Hengel, and Qi Wu. Object-and-action aware model for visual language navigation. In *European Conference on Computer Vision*, pp. 303–317. Springer, 2020a.

Yuankai Qi, Qi Wu, Peter Anderson, Xin Wang, William Yang Wang, Chunhua Shen, and Anton van den Hengel. Reverie: Remote embodied visual referring expression in real indoor environments. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9982–9991, 2020b.

Yuankai Qi, Zizheng Pan, Yicong Hong, Ming-Hsuan Yang, Anton van den Hengel, and Qi Wu. 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*, pp. 1655–1664, 2021.

Liu Qianyi-Liu Jiajun Liu Jing Wu Qi Qiao, Yanyuan. Llm as copilot for coarse-grained vision-and-language navigation. *ECCV*, 2024.

Yanyuan Qiao, Yuankai Qi, Yicong Hong, Zheng Yu, Peng Wang, and Qi Wu. HOP: history-and-order aware pre-training for vision-and-language navigation. In *Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition*, pp. 8524–8537, 2022.

Yanyuan Qiao, Yuankai Qi, Yicong Hong, Zheng Yu, Peng Wang, and Qi Wu. Hop+: History-enhanced and order-aware pre-training for vision-and-language navigation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023a.

Yanyuan Qiao, Yuankai Qi, Zheng Yu, Jing Liu, and Qi Wu. March in chat: Interactive prompting for remote embodied referring expression. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 15758–15767, 2023b.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. pp. 8748–8763, 2021.

Abhinav Rajvanshi, Karan Sikka, Xiao Lin, Bhoram Lee, Han-Pang Chiu, and Alvaro Velasquez. Saynav: Grounding large language models for dynamic planning to navigation in new environments. *arXiv preprint arXiv:2309.04077*, 2023.

Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *International conference on machine learning*, pp. 8821–8831. Pmlr, 2021.

Krishan Rana, Jesse Haviland, Sourav Garg, Jad Abou-Chakra, Ian Reid, and Niko Suenderhauf. Sayplan: Grounding large language models using 3d scene graphs for scalable robot task planning. In *7th Annual Conference on Robot Learning*, 2023.

Sonia Raychaudhuri, Saim Wani, Shivansh Patel, Unnat Jain, and Angel X. Chang. Language-aligned waypoint (LAW) supervision for vision-and-language navigation in continuous environments. In *EMNLP*, pp. 4018–4028, 2021.

Allen Z Ren, Anushri Dixit, Alexandra Bodrova, Sumeet Singh, Stephen Tu, Noah Brown, Peng Xu, Leila Takayama, Fei Xia, Jake Varley, et al. Robots that ask for help: Uncertainty alignment for large language model planners. *Proceedings of Machine Learning Research*, 229, 2023.

T Rodrigo. Navigational strategies and models. *Psicológica*, 23(1), 2002.

Junha Roh, Chris Paxton, Andrzej Pronobis, Ali Farhadi, and Dieter Fox. Conditional driving from natural language instructions. In *Proceedings of the Conference on Robot Learning*, pp. 540–551, 2020.

Homero Roman Roman, Yonatan Bisk, Jesse Thomason, Asli Celikyilmaz, and Jianfeng Gao. Rmm: A recursive mental model for dialogue navigation. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pp. 1732–1745, 2020.

Homagni Saha, Fateme Fotouhi, Qisai Liu, and Soumik Sarkar. A modular vision language navigation and manipulation framework for long horizon compositional tasks in indoor environment. *Frontiers in Robotics and AI*, 9, 2022.

Gabriel Sarch, Yue Wu, Michael J Tarr, and Katerina Fragkiadaki. Open-ended instructable embodied agents with memory-augmented large language models. *arXiv preprint arXiv:2310.15127*, 2023.

Manolis Savva, Jitendra Malik, Devi Parikh, Dhruv Batra, Abhishek Kadian, Oleksandr Maksymets, Yili Zhao, Erik Wijmans, Bhavana Jain, Julian Straub, Jia Liu, and Vladlen Koltun. Habitat: A platform for embodied AI research. pp. 9338–9346, 2019.

Hao Sha, Yao Mu, Yuxuan Jiang, Li Chen, Chenfeng Xu, Ping Luo, Shengbo Eben Li, Masayoshi Tomizuka, Wei Zhan, and Mingyu Ding. Languagempc: Large language models as decision makers for autonomous driving. *arXiv preprint arXiv:2310.03026*, 2023.

Dhruv Shah, Błażej Osiński, Sergey Levine, et al. Lm-nav: Robotic navigation with large pre-trained models of language, vision, and action. In *Conference on robot learning*, pp. 492–504. PMLR, 2023.

Hao Shao, Yuxuan Hu, Letian Wang, Steven L Waslander, Yu Liu, and Hongsheng Li. Lmdrive: Closed-loop end-to-end driving with large language models. *arXiv preprint arXiv:2312.07488*, 2023.

Sheng Shen, Liunian Harold Li, Hao Tan, Mohit Bansal, Anna Rohrbach, Kai-Wei Chang, Zhewei Yao, and Kurt Keutzer. How much can clip benefit vision-and-language tasks? *arXiv preprint arXiv:2107.06383*, 2021.

Mohit Shridhar, Jesse Thomason, Daniel Gordon, Yonatan Bisk, Winson Han, Roozbeh Mottaghi, Luke Zettlemoyer, and Dieter Fox. Alfred: A benchmark for interpreting grounded instructions for everyday tasks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10740–10749, 2020.

Anna Shusterman, Sang Ah Lee, and Elizabeth S Spelke. Cognitive effects of language on human navigation. *Cognition*, 120(2):186–201, 2011.

Chonghao Sima, Katrin Renz, Kashyap Chitta, Li Chen, Hanxue Zhang, Chengen Xie, Ping Luo, Andreas Geiger, and Hongyang Li. Drivelm: Driving with graph visual question answering. *arXiv preprint arXiv:2312.14150*, 2023.

Kunal Pratap Singh, Luca Weihs, Alvaro Herrasti, Jonghyun Choi, Aniruddha Kembhavi, and Roozbeh Mottaghi. Ask4help: Learning to leverage an expert for embodied tasks. *Advances in Neural Information Processing Systems*, 35:16221–16232, 2022.

Chan Hee Song, Jiaman Wu, Clayton Washington, Brian M Sadler, Wei-Lun Chao, and Yu Su. Llm-planner: Few-shot grounded planning for embodied agents with large language models. *arXiv preprint arXiv:2212.04088*, 2022.

NN Sriram, Tirth Maniar, Jayaganesh Kalyanasundaram, Vineet Gandhi, Brojeshwar Bhowmick, and K Madhava Krishna. Talk to the vehicle: Language conditioned autonomous navigation of self driving cars. In *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 5284–5290. IEEE, 2019.

Yifei Su, Dong An, Yuan Xu, Kehan Chen, and Yan Huang. Target-grounded graph-aware transformer for aerial vision-and-dialog navigation. *arXiv preprint arXiv:2308.11561*, 2023.

Andrew Szot, Alexander Clegg, Eric Undersander, Erik Wijmans, Yili Zhao, John M. Turner, Noah Maestre, Mustafa Mukadam, Devendra Singh Chaplot, Oleksandr Maksymets, Aaron Gokaslan, Vladimir Vondrus, Sameer Dharur, Franziska Meier, Wojciech Galuba, Angel X. Chang, Zsolt Kira, Vladlen Koltun, Jitendra Malik, Manolis Savva, and Dhruv Batra. Habitat 2.0: Training home assistants to rearrange their habitat. In *Advances in Neural Information Processing Systems*, pp. 251–266, 2021.

Hao Tan and Mohit Bansal. Lxmert: Learning cross-modality encoder representations from transformers. *arXiv preprint arXiv:1908.07490*, 2019.

Hao Tan, Licheng Yu, and Mohit Bansal. Learning to navigate unseen environments: Back translation with environmental dropout. *arXiv preprint arXiv:1904.04195*, 2019.

Jesse Thomason, Michael Murray, Maya Cakmak, and Luke Zettlemoyer. Vision-and-dialog navigation. In *Conference on Robot Learning*, pp. 394–406. PMLR, 2020.

Xiaoyu Tian, Junru Gu, Bailin Li, Yicheng Liu, Chenxu Hu, Yang Wang, Kun Zhan, Peng Jia, Xianpeng Lang, and Hang Zhao. Drivevlm: The convergence of autonomous driving and large vision-language models. *arXiv preprint arXiv:2402.12289*, 2024.

Edward C Tolman. Cognitive maps in rats and men. *Psychological review*, 55(4):189, 1948.

Dennis Ulmer, Elman Mansimov, Kaixiang Lin, Justin Sun, Xibin Gao, and Yi Zhang. Bootstrapping llm-based task-oriented dialogue agents via self-talk. *arXiv preprint arXiv:2401.05033*, 2024.

Karthik Valmeekam, Alberto Olmo, Sarath Sreedharan, and Subbarao Kambhampati. Planbench: An extensible benchmark for evaluating large language models on planning and reasoning about change. In *Neural Information Processing Systems*, 2022.

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

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

Hanqing Wang, Wei Liang, Luc Van Gool, and Wenguan Wang. Dreamwalker: Mental planning for continuous vision-language navigation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 10873–10883, 2023a.

Liuyi Wang, Zongtao He, Jiagui Tang, Ronghao Dang, Naijia Wang, Chengju Liu, and Qijun Chen. A dual semantic-aware recurrent global-adaptive network for vision-and-language navigation. *arXiv preprint arXiv:2305.03602*, 2023b.

Liuyi Wang, Chengju Liu, Zongtao He, Shu Li, Qingqing Yan, Huiyi Chen, and Qi Chen. Pasts: Progress-aware spatio-temporal transformer speaker for vision-and-language navigation. *Eng. Appl. Artif. Intell.*, 128:107487, 2023c. URL `https://api.semanticscholar.org/CorpusID:258833536`.

Su Wang, Ceslee Montgomery, Jordi Orbay, Vighnesh Birodkar, Aleksandra Faust, Izzeddin Gur, Natasha Jaques, Austin Waters, Jason Baldridge, and Peter Anderson. Less is more: Generating grounded navigation instructions from landmarks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15428–15438, 2022a.

Wenhai Wang, Jiangwei Xie, ChuanYang Hu, Haoming Zou, Jianan Fan, Wenwen Tong, Yang Wen, Silei Wu, Hanming Deng, et al. Drivemlm: Aligning multi-modal large language models with behavioral planning states for autonomous driving. *arXiv preprint arXiv:2312.09245*, 2023d.

Xiaohan Wang, Wenguan Wang, Jiayi Shao, and Yi Yang. Lana: A language-capable navigator for instruction following and generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 19048–19058, 2023e.

Xiaohan Wang, Wenguan Wang, Jiayi Shao, and Yi Yang. Learning to follow and generate instructions for language-capable navigation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023f.

Xin Wang, Qiuyuan Huang, Asli Celikyilmaz, Jianfeng Gao, Dinghan Shen, Yuan-Fang Wang, William Yang Wang, and Lei Zhang. Reinforced cross-modal matching and self-supervised imitation learning for vision-language navigation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 6629–6638, 2019.

Yi Wang, Kunchang Li, Yizhuo Li, Yinan He, Bingkun Huang, Zhiyu Zhao, Hongjie Zhang, Jilan Xu, Yi Liu, Zun Wang, et al. Internvideo: General video foundation models via generative and discriminative learning. *arXiv preprint arXiv:2212.03191*, 2022b.

Zehao Wang, Mingxiao Li, Minye Wu, Marie-Francine Moens, and Tinne Tuytelaars. Find a way forward: a language-guided semantic map navigator. *arXiv preprint arXiv:2203.03183*, 2022c.

Zihan Wang, Xiangyang Li, Jiahao Yang, Yeqi Liu, and Shuqiang Jiang. Gridmm: Grid memory map for vision-and-language navigation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 15625–15636, 2023g.

Zihan Wang, Xiangyang Li, Jiahao Yang, Yeqi Liu, Junjie Hu, Ming Jiang, and Shuqiang Jiang. Lookahead exploration with neural radiance representation for continuous vision-language navigation. *arXiv preprint arXiv:2404.01943*, 2024a.

Zihan Wang, Xiangyang Li, Jiahao Yang, Yeqi Liu, and Shuqiang Jiang. Sim-to-real transfer via 3d feature fields for vision-and-language navigation. *arXiv preprint arXiv:2406.09798*, 2024b.

Zun Wang, Jialu Li, Yicong Hong, Yi Wang, Qi Wu, Mohit Bansal, Stephen Gould, Hao Tan, and Yu Qiao. Scaling data generation in vision-and-language navigation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 12009–12020, 2023h.

William H Warren. Non-euclidean navigation. *Journal of Experimental Biology*, 222(Suppl_1):jeb187971, 2019.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*, 2022.

Licheng Wen, Daocheng Fu, Xin Li, Xinyu Cai, Tao Ma, Pinlong Cai, Min Dou, Botian Shi, Liang He, and Yu Qiao. Dilu: A knowledge-driven approach to autonomous driving with large language models. *arXiv preprint arXiv:2309.16292*, 2023.

Wansen Wu, Tao Chang, Xinmeng Li, Quanjun Yin, and Yue Hu. Vision-language navigation: a survey and taxonomy. *Neural Computing and Applications*, 36(7):3291–3316, 2024.

Zongkai Wu, Zihan Liu, and Donglin Wang. Multi-grounding navigator for self-supervised vision-and-language navigation. In *2021 International Joint Conference on Neural Networks (IJCNN)*, pp. 1–8. IEEE, 2021.

Fei Xia, Amir R Zamir, Zhiyang He, Alexander Sax, Jitendra Malik, and Silvio Savarese. Gibson env: Real-world perception for embodied agents. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 9068–9079, 2018.

Qiaolin Xia, Xiujun Li, Chunyuan Li, Yonatan Bisk, Zhifang Sui, Jianfeng Gao, Yejin Choi, and Noah A Smith. Multi-view learning for vision-and-language navigation. *arXiv preprint arXiv:2003.00857*, 2020.

Jiannan Xiang, Tianhua Tao, Yi Gu, Tianmin Shu, Zirui Wang, Zichao Yang, and Zhiting Hu. Language models meet world models: Embodied experiences enhance language models. *Advances in neural information processing systems*, 36, 2024.

Ning Xie, Farley Lai, Derek Doran, and Asim Kadav. Visual entailment: A novel task for fine-grained image understanding. *arXiv preprint arXiv:1901.06706*, 2019.

Zhenhua Xu, Yujia Zhang, Enze Xie, Zhen Zhao, Yong Guo, Kenneth KY Wong, Zhenguo Li, and Hengshuang Zhao. Drivegpt4: Interpretable end-to-end autonomous driving via large language model. *arXiv preprint arXiv:2310.01412*, 2023.

Zhiyuan Xu, Kun Wu, Junjie Wen, Jinming Li, Ning Liu, Zhengping Che, and Jian Tang. A survey on robotics with foundation models: toward embodied ai. *arXiv preprint arXiv:2402.02385*, 2024.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. mt5: A massively multilingual pre-trained text-to-text transformer. *arXiv preprint arXiv:2010.11934*, 2020.

Xu Yan, Haiming Zhang, Yingjie Cai, Jingming Guo, Weichao Qiu, Bin Gao, Kaiqiang Zhou, Yue Zhao, Huan Jin, Jiantao Gao, et al. Forging vision foundation models for autonomous driving: Challenges, methodologies, and opportunities. *arXiv preprint arXiv:2401.08045*, 2024.

Jianing Yang, Xuweiyi Chen, Shengyi Qian, Nikhil Madaan, Madhavan Iyengar, David F. Fouhey, and Joyce Chai. Llm-grounder: Open-vocabulary 3d visual grounding with large language model as an agent. In *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, 2024.

Jingkang Yang, Yuhao Dong, Shuai Liu, Bo Li, Ziyue Wang, Chencheng Jiang, Haoran Tan, Jiamu Kang, Yuanhan Zhang, Kaiyang Zhou, et al. Octopus: Embodied vision-language programmer from environmental feedback. *arXiv preprint arXiv:2310.08588*, 2023.

Qinghao Ye, Haiyang Xu, Guohai Xu, Jiabo Ye, Ming Yan, Yiyang Zhou, Junyang Wang, Anwen Hu, Pengcheng Shi, Yaya Shi, et al. mplug-owl: Modularization empowers large language models with multimodality. *arXiv preprint arXiv:2304.14178*, 2023.

Sriram Yenamandra, Arun Ramachandran, Karmesh Yadav, Austin Wang, Mukul Khanna, Theophile Gervet, Tsung-Yen Yang, Vidhi Jain, Alexander William Clegg, John Turner, Zsolt Kira, Manolis Savva, Angel Chang, Devendra Singh Chaplot, Dhruv Batra, Roozbeh Mottaghi, Yonatan Bisk, and Chris Paxton. Homerobot: Open-vocabulary mobile manipulation, 2024. URL `https://arxiv.org/abs/2306.11565`.

Jianhao Yuan, Shuyang Sun, Daniel Omeiza, Bo Zhao, Paul Newman, Lars Kunze, and Matthew Gadd. Rag-driver: Generalisable driving explanations with retrieval-augmented in-context learning in multi-modal large language model. *arXiv preprint arXiv:2402.10828*, 2024.

Yuexiang Zhai, Hao Bai, Zipeng Lin, Jiayi Pan, Shengbang Tong, Yifei Zhou, Alane Suhr, Saining Xie, Yann LeCun, Yi Ma, et al. Fine-tuning large vision-language models as decision-making agents via reinforcement learning. *arXiv preprint arXiv:2405.10292*, 2024.

Zhaohuan Zhan, Jinghui Qin, Wei Zhuo, and Guang Tan. Enhancing vision and language navigation with prompt-based scene knowledge. *IEEE Transactions on Circuits and Systems for Video Technology*, 2024a.

Zhaohuan Zhan, Lisha Yu, Sijie Yu, and Guang Tan. Mc-gpt: Empowering vision-and-language navigation with memory map and reasoning chains. *arXiv preprint arXiv:2405.10620*, 2024b.

Jiazhao Zhang, Kunyu Wang, Rongtao Xu, Gengze Zhou, Yicong Hong, Xiaomeng Fang, Qi Wu, Zhizheng Zhang, and Wang He. Navid: Video-based vlm plans the next step for vision-and-language navigation. In *Robotics: Science and Systems (RSS)*, 2024a.

Tianyao Zhang, Xiaoguang Hu, Jin Xiao, and Guofeng Zhang. A survey of visual navigation: From geometry to embodied ai. *Engineering Applications of Artificial Intelligence*, 114:105036, 2022a.

Yichi Zhang, Jianing Yang, Jiayi Pan, Shane Storks, Nikhil Devraj, Ziqiao Ma, Keunwoo Peter Yu, Yuwei Bao, and Joyce Chai. Danli: Deliberative agent for following natural language instructions. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 2022b.

Yichi Zhang, Jianing Yang, Keunwoo Yu, Yinpei Dai, Shane Storks, Yuwei Bao, Jiayi Pan, Nikhil Devraj, Ziqiao Ma, and Joyce Chai. Seagull: An embodied agent for instruction following through situated dialog. In *Alexa Prize SimBot Challenge Proceedings*, 2023.

Yubo Zhang, Hao Tan, and Mohit Bansal. Diagnosing the environment bias in vision-and-language navigation. *arXiv preprint arXiv:2005.03086*, 2020.

Yue Zhang and Parisa Kordjamshidi. Lovis: Learning orientation and visual signals for vision and language navigation. pp. 5745–5754, 2022a.

Yue Zhang and Parisa Kordjamshidi. Explicit object relation alignment for vision and language navigation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pp. 322–331, 2022b.

Yue Zhang and Parisa Kordjamshidi. Vln-trans: Translator for the vision and language navigation agent. *arXiv preprint arXiv:2302.09230*, 2023.

Yue Zhang, Quan Guo, and Parisa Kordjamshidi. Towards navigation by reasoning over spatial configurations. *arXiv preprint arXiv:2105.06839*, 2021.

Yue Zhang, Quan Guo, and Parisa Kordjamshidi. Navhint: Vision and language navigation agent with a hint generator. *arXiv preprint arXiv:2402.02559*, 2024b.

Ming Zhao, Peter Anderson, Vihan Jain, Su Wang, Alexander Ku, Jason Baldridge, and Eugene Ie. On the evaluation of vision-and-language navigation instructions. *arXiv preprint arXiv:2101.10504*, 2021.

Yusheng Zhao, Jinyu Chen, Chen Gao, Wenguan Wang, Lirong Yang, Haibing Ren, Huaxia Xia, and Si Liu. Target-driven structured transformer planner for vision-language navigation. pp. 4194–4203, 2022.

Duo Zheng, Shijia Huang, Lin Zhao, Yiwu Zhong, and Liwei Wang. Towards learning a generalist model for embodied navigation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13624–13634, 2024.

Kaizhi Zheng, Kaiwen Zhou, Jing Gu, Yue Fan, Jialu Wang, Zonglin Di, Xuehai He, and Xin Eric Wang. Jarvis: A neuro-symbolic commonsense reasoning framework for conversational embodied agents. *arXiv preprint arXiv:2208.13266*, 2022.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric. P Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging llm-as-a-judge with mt-bench and chatbot arena, 2023a.

Qi Zheng, Daqing Liu, Chaoyue Wang, Jing Zhang, Dadong Wang, and Dacheng Tao. Esceme: Vision-and-language navigation with episodic scene memory. *arXiv preprint arXiv:2303.01032*, 2023b.

Ce Zhou, Qian Li, Chen Li, Jun Yu, Yixin Liu, Guangjing Wang, Kai Zhang, Cheng Ji, Qiben Yan, Lifang He, et al. A comprehensive survey on pretrained foundation models: A history from bert to chatgpt. *arXiv preprint arXiv:2302.09419*, 2023.

Gengze Zhou, Yicong Hong, and Qi Wu. Navgpt: Explicit reasoning in vision-and-language navigation with large language models. pp. 7641–7649. AAAI Press, 2024a. doi: 10.1609/AAAI.V38I7.28597. URL `https://doi.org/10.1609/aaai.v38i7.28597`.

Gengze Zhou, Yicong Hong, and Qi Wu. Navgpt: Explicit reasoning in vision-and-language navigation with large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 7641–7649, 2024b.

Qinhong Zhou, Sunli Chen, Yisong Wang, Haozhe Xu, Weihua Du, Hongxin Zhang, Yilun Du, Joshua B Tenenbaum, and Chuang Gan. Hazard challenge: Embodied decision making in dynamically changing environments. In *The Twelfth International Conference on Learning Representations*, 2024c.

Xinzhe Zhou and Yadong Mu. Tree-structured trajectory encoding for vision-and-language navigation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 3814–3824, 2023.

Chen Zhu, Michael Meurer, and Christoph Günther. Integrity of visual navigation—developments, challenges, and prospects. *NAVIGATION: Journal of the Institute of Navigation*, 69(2), 2022.

Fengda Zhu, Yi Zhu, Xiaojun Chang, and Xiaodan Liang. Vision-language navigation with self-supervised auxiliary reasoning tasks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10012–10022, 2020.

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

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

Fengda Zhu, Vincent CS Lee, Xiaojun Chang, and Xiaodan Liang. Vision language navigation with knowledge-driven environmental dreamer. In *International Joint Conference on Artificial Intelligence 2023*, pp. 1840–1848. Association for the Advancement of Artificial Intelligence (AAAI), 2023.

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