Reason & Act : A Modular Approach to Explanation Driven Agents for Vision and Language Navigation

Shaunak Halbe[1], Ingrid Navarro[2] and Jean Oh[2]

***Abstract*— Vision-and-Language Navigation (VLN) is a multimodal task where an agent follows natural language instructions to navigate in photo-realistic environments. VLN assumes discrete motion along viewpoints of an undirected navigation graph. However, navigation in the real world demands continuous movement through low-level actions, thus motivating the task of Vision-and-Language Navigation in Continuous Environments (VLN-CE). Current approaches to VLN-CE use end-to-end models that attempt to solve both global reasoning and low-level control tasks. Training a single model to perform tasks with vastly differing requirements is difficult. We present the design of a modular system in the form of a global and local planner. The global planner would be responsible for the overall navigation to the desired goal position as indicated by the natural language instruction. It predicts a high-level waypoint to be reached by a local planner through execution of a series of low-level actions. The current baselines for VLN-CE are weak and cannot be scaled for global planning. In this paper, we focus on improving multi-modal understanding of VLN-CE agents with an intention of extending them to form the global planner. To boost multi-modal understanding, we introduce a grounding module along with a Reason-and-Act strategy requiring the agent to identify salient objects in its surroundings. Such a scheme allows the agent to derive visual cues and match them with the verbal indicators given in the instruction. We believe, an agent that can learn to link the signals present in different modalities can perform better in unseen environments.**

***Index Terms*— Vision-and-Language Navigation, Embodied Agents, Hierarchical Planning**

I. INTRODUCTION

A robot that can understand and execute human instructions has been a dream for scientists since ages. Up until a few years ago, such a robot was only imagined in science fiction movies. Vision-and-Language Navigation (VLN) [1] takes a significant step towards achieving this dream by formally defining this task. VLN requires an agent to navigate across photo-realistic visual scenes by inferring directional cues from a natural language instruction. Although this is an inherently challenging task for robots to carry out, certain assumptions have simplified the requirements for developing such an agent. VLN agents move by snapping across discrete viewpoints of an undirected navigation graph and are not concerned with the low-level path planning required to reach any viewpoint. These agents observe the environment through panoramic images, and use it to choose the next viewpoint from a list of available candidates.

Some of these assumptions are strong as compared to real world conditions. The more recently proposed task of Vision-and-Language Navigation in Continuous Environments (VLN-CE) [2] takes a step closer to the real world setting, by requiring the agents to execute low-level actions in continuous environments. This setting presents further challenges as the agents are no longer guaranteed perfect localization, actuation, and navigation. The authors of VLN-CE [2] introduce two end-to-end models to serve as baselines. Due to the complex nature of this task, the models achieve low success rates.

We believe that solving such a complex, multi-stage task requires a hierarchical approach with modular components that divide task responsibilities (e.g., alignment, reasoning, control, etc) among themselves. Toward this end, we explore methods for improving the high-level planning aspect. Specifically, we focus on improving the alignment between visual and verbal signals with a goal of leveraging it to improve high-level navigation.

We discuss the structure of a global planner which is entrusted with the task of correlating the visual observations with the instruction and providing us with a high level waypoint to navigate towards. Such a waypoint would then be reached by a local planner through the execution of a series of low-level actions. We explore the idea of an agent that can identify salient features in visual scenes and link them to verbal indicators to develop a richer understanding of the environment. In this spirit, we introduce a reasoning component, which requires the model to identify salient objects in its surroundings that are pivotal in navigating towards the goal. To summarize, our contributions to improve the high-level planning are two fold; we

- introduce a Vision-Language grounding module that generates strongly grounded features in Vision, Depth and Language Space.
- propose a reasoning component that allows an agent to enhance its multi-modal understanding.

II. RELATED WORK

A. Vision-Language Navigation

In VLN [1], an agent is required to follow a navigation instruction from a start location to a goal. Usually, the goal position is not explicitly provided and is to be inferred from the instruction. Overall, VLN models ([3]–[6]) have seen considerable progress in their ability to reach the goal and

[1] S. Halbe is with the Department of Computer Engineering, College of Engineering, Pune, Maharashtra, India. halbesa18.comp@coep.ac.in

[2] I. Navarro and J. Oh are with the Robotics Institute, Carnegie Mellon University, Pittsburgh, PA., USA ingridn@cs.cmu.edu, jeanoh@nrec.ri.cmu.edu

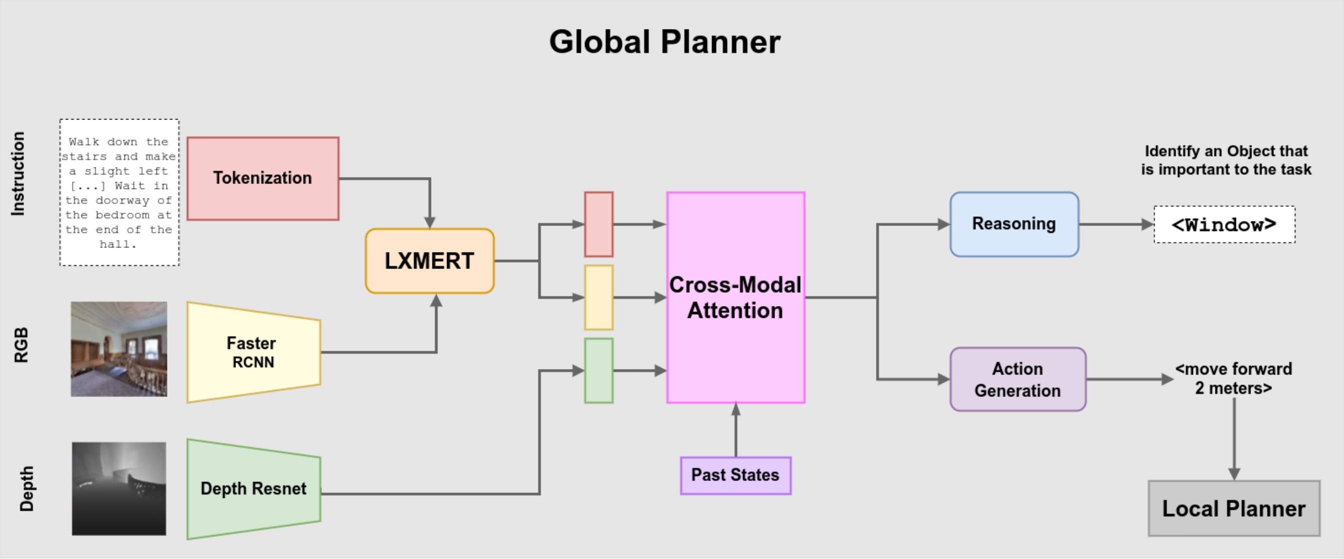

Fig. 1. Global Planner with Cross Modal Attention

the degree of their instruction-trajectory alignment. While most existing works only consider the scenario where the test environments are previously unexplored, some ([3], [5], [7]) also consider a setting where the agent can explore the test scenes prior to evaluation. Among these works, the *Speaker-Follower* [3] approach is quite common where a *speaker* model generates novel instructions from sampled trajectories which are then used for path selection while training the follower. Different from these, transformer based Vision-and-Language pre-training approaches have been successfully extended to VLN. [6], [8] have achieved positive results by demonstrably increasing the vision and language alignment by transferring multimodal transformers pretrained on internet data to VLN settings.

B. *Other Language-Guided Navigation Tasks*

Apart from VLN, several other tasks involving Language-Guided Navigation / Interaction ([2], [9]–[13]) have been proposed which place an agent in an embodied setting requiring visio-linguistic understanding. These tasks are in a similar vein to the VLN tasks but differ in the activity expected from the agent. Most similar to VLN is VLN-CE [2], which requires an agent to move in a continuous environment in the absence of a navigation graph. VLN-CE also differs with regard to the topological and positional knowledge that the agent has access to. However, VLN-CE has the same high-level objective of language guided navigation as VLN. On the other hand, Embodied Question Answering (EQA) [9] requires an agent to navigate based on the natural language question and answer it using the explored information. Similarly, in Embodied Object Referral (EOR) the agent is tasked with navigating towards an object in the environment based on a natural language instruction. Unlike EOR and EQA, tasks like Vision and Dialog History Navigation (VDHN) and Embodied Goal-Directed Manipulation (EGM) require interaction with the oracle or manipulation in the environment. We refer readers to [14] for further details about the aforementioned tasks. In this paper we focus on designing a modular agent for the VLN-CE task.

C. *Modular Planning*

Previous works([12], [15], [16]) have proposed hierarchical approaches to solve Embodied Vision-and-Language Planning tasks. [17] propose *MoViLan*, a modular approach for long horizon tasks such as Vision-and-Language Navigation. *MoViLan* uses a novel Graph Convolutional Neural Network (G-CNN) based approach for mapping by approximating the geometry of nearby objects. The navigation map thus generated is used along with semantic information to predict high-level actions. Finally these high level actions are decomposed into low-level actions using a non-learning search strategy like $A*$. [15] and [16] use supervised learning to learn to predict high-level waypoints using images and instructions. In the second stage, Reinforcement Learning is used to learn actions to reach these waypoints.

Similar to these approaches, we discuss a modular design to optimize for subgoals using a global planner. However, in VLN-CE environment subgoals are not explicitly provided making it a challenging task to work on.

III. PROBLEM FORMULATION

Following the definition in [14], let $\mathcal{S} = \{\mathcal{V}, \mathcal{L}\}$ represent the set of states encompassing the visual observations, $\mathcal{V}$, and language inputs, $\mathcal{L}$. Next, let $\mathcal{A} = \{\mathtt{stop}, \mathtt{turn\_left}, \mathtt{turn\_right}, \mathtt{move\_forward}\}$ include the set of possible actions. The VLN task can be formulated as $\Phi_{\text{VLN}} = \{\mathcal{S}, \mathcal{A}, s_0, s_{\text{goal}}\}$, where $s_0, s_{\text{goal}} \in \mathcal{S}$ represent the initial and target states, respectively. Thus, a plan $\Psi_{\text{VLN}} = \langle s_0, a_0, s_1, a_1, \ldots, s_{\text{T}}, a_{\text{T}} \rangle$ exists such that each state s_{t}, where $t \in [0, T]$, is associated with a location in the environment leading to the final goal. An episode in

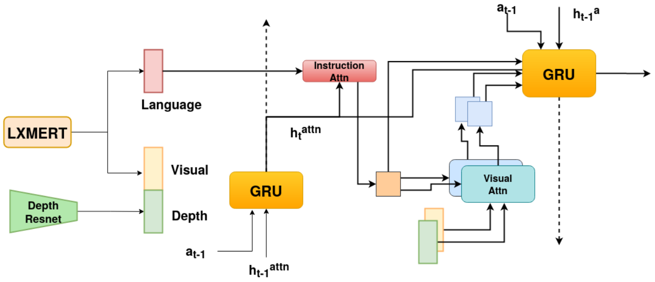

Fig. 2. Cross Modal Attention

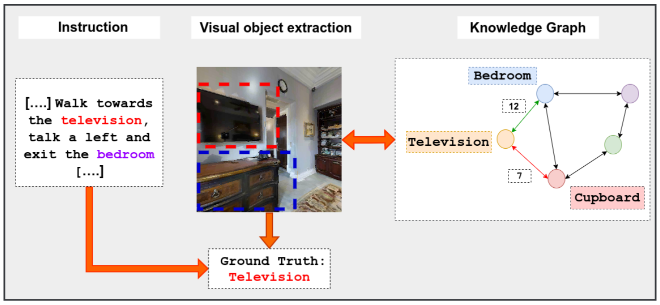

Fig. 3. Extracting the ground truth from Scene Priors

VLN requires an agent to find a route from the start state to a target state following an instruction $l \in \mathcal{L}$. At each time-step t, the agent in the environment $\mathcal{E}$ is said to be in a state $\mathrm{s_t}$, represented as ($\mathrm{v_t}$, l) where $\mathrm{v_t}$ corresponds to the current visual observation, and l is the instruction. The agent must predict a solution $\widehat{\Psi}_{\mathrm{VLN}} = \langle \mathrm{s_0,a_0,s_1,a_1,...,s_T,a_T} \rangle$ by executing an action $\mathrm{a_t} \in \mathcal{A}$ at each state $\mathrm{s_t}$ following a policy π parametrized by θ such that $a_{\mathrm{t}} = \pi(s_t; \theta)$.

The episode is deemed successful if the sequence of actions, both, delivers the agent close to the intended goal location $\mathrm{s_{goal}}$, and minimizes the difference between the ground-truth plan Ψ_{VLN}, and the predicted plan $\widehat{\Psi}_{\mathrm{VLN}}$.

IV. APPROACH

In this section, we introduce our global planner π_{global} which is tasked with predicting the next waypoint given visual observations and a specified instruction. We assume the predicted waypoints are passed to a local planner π_{local} which predicts the sequence of low-level actions to reach each of the intended waypoints. As mentioned in the previous sections, our paper focuses on exploring techniques to improve the high-level planning. Thus, we leave local planning out of the scope of our work. We refer readers to Figure 1 for the model architecture.

A. Global Planner

Following [2], we leverage imitation learning [18] to train the global policy π_{global} to predict the next waypoint g_{t} by imitating expert actions. We train our global agent using AI Habitat [19] and the VLN-CE dataset [2]. In our setting, π_{global} receives an instruction and at each time-step has access to visual observations comprised by color and depth images. The global policy then uses this information to predict the next waypoint $g_{\mathrm{t}} = \pi_{\mathrm{global}}(s_t; \theta_{\mathrm{global}})$ in terms of distance and heading relative to its current position.

Our global planner is comprised by two sub-modules, a grounding module tasked with ensuring the alignment between the language and visual modalities, and a reasoning component which ensures the agent is able to explain the actions taken in the past. Below we provide further details about the aforementioned modules.

1) Grounding Module: In VLN-CE [2], the authors propose a vanilla Sequence-to-Sequence (Seq2Seq) model and a Cross-Modal Attention (CMA) based Recurrent Neural Network (RNN) to serve as baselines for the tasks. Pre-trained transformer models ([20]–[22]) have starkly outperformed RNN based approaches across a range of language only (Question-Answering, Language Modelling) and multimodal (VQA [23], VisDial [24]) tasks. Drawing inspiration from such tasks, we introduce a Vision-and-Language Grounding module in the form of a pre-trained LXMERT encoder [25]. This module completely replaces the individual instruction encoder and RGB image encoder from the baselines. By design, said model requires bounding box feature vectors of top 36 objects extracted by a 101-layer Faster-RCNN. Thus, we use a pre-trained Faster-RCNN [26] model to extract objects features from our RGB observations to feed to the transformer-based model. We freeze the parameters of both of the encoders, and merely use them as feature extractors. The LXMERT model encodes the image features and instruction tokens and performs cross-modal as well as self-attention over 9 language, 5 visual and 5 cross-modal transformer-encoder layers. For each image-instruction pair, we extract the last layer's outputs from language and vision streams of LXMERT and combine it to form a representation grounded in vision and language. Separately, we use a Resnet [27] encoder trained on the dataset from the Gibson Environment [28] to extract features from depth observations.

Finally, we use CMA as in [2] to fuse the grounded features extracted from LXMERT with the depth features. CMA consists of two RNN encoders as shown in Figure 2, one to track visual observations and the other one to make decisions based on attended features. The previous action features along with the hidden state are used to attend over the language embedding from LXMERT. This attended language embedding is in-turn used to attend to the visual and depth features. Thus through cross-modal interaction, a strongly grounded representation is produced.

2) Reasoning Component: Through this component, we task the agent with identifying an object in its field of sight that is most relevant to instruction and the region that the agent is in. We implement the reasoning module as a linear layer on top of the attention module. We pass the grounded features through the linear layer with an aim to classify it between the 41 object categories present in Matterport3D [29]. We use Cross Entropy as the loss function and optimize it auxiliary to the action prediction loss. Curating good-quality ground truths for each scene is very crucial. At each step, we choose the ground truth object in one of following

	Val-Seen							Val-Unseen						
	PL ↓	**NE ↓**	**nDTW ↑**	**OS ↑**	**SR ↑**	**SPL ↑**	**ST ↓**	**PL ↓**	**NE ↓**	**nDTW ↑**	**OS ↑**	**SR ↑**	**SPL ↑**	**ST ↓**
Seq2Seq w/o reasoning	**7.60**	**8.48**	**45.60**	30.20	23.52	**22.41**	102	7.77	9.14	40.73	**25.28**	**16.53**	15.03	97
Seq2Seq w/ reasoning (Ours)	8.30	8.66	44.95	**34.83**	**23.65**	21.99	**99**	**7.50**	**8.88**	**42.40**	23.54	16.47	**15.28**	**87**

TABLE I

REASONING EXPERIMENT

	Val-Unseen						
	PL ↓	**NE ↓**	**nDTW ↑**	**OS ↑**	**SR ↑**	**SPL ↑**	**ST ↓**
CMA	8.59	9.20	41.49	**28.16**	**17.45**	15.82	114
LXMERT + CMA (Ours)	**8.31**	**9.02**	**42.21**	27.19	17.18	**15.92**	**100**

TABLE II

ALIGNMENT EXPERIMENT

three ways:

- An object directly mentioned in the instruction is present in the agent's field of sight
- An object present in the visual scene is correlated to an object mentioned in the instruction
- An object present in the visual scene is often observed in the region (room) where the agent is currently located

We use a Knowledge Graph from Visual Genome[30] to find associations between objects and determine their co-occurrence. We filter this Knowledge Graph by keeping only the objects and regions present in Matterport. Given an object, we use the Knowledge Graph to find other commonly associated objects. While choosing the ground truth an object directly mentioned in the instruction is given highest preference. In case of multiple objects, we use co-occurrence values to determine the ground truth. At each step we maintain a list of objects consisting of the ones directly mentioned in the instructions, associated with the objects mentioned in the instruction and ones that are associated with the region (room) where the agent is present. We extract co-occurrence values between two objects and between an object and a region from the Knowledge Graph. We select the object with the maximum co-occurrence value as the ground truth for the reasoning task.

V. EXPERIMENTS

A. Metrics

We report standard metrics for visual navigation tasks defined in [1], [31], [32] of success rate (SR), success weighted by inverse path length (SPL), normalized dynamic-time warping (nDTW), path length in meters (PL), oracle success rate (OS), navigation error in meters from goal at termination (NE), and steps taken (ST) to quantify the performance of the model.

B. Implementation Details

We train our agents on the 'train' split from VLN-CE dataset in the AI Habitat simulator[19]. We utilize the Adam optimizer [33] with a learning rate of 2.5×10^{-4}. We use a DAgger-like [18] approach to collect trajectories with oracle actions as ground truth actions. We collect 10,819 trajectories for both of the experiments. Imitation learning is then performed for 15 epochs over all collected trajectories. In order to match the original setup [2], we set the forward actuation of the agent to 0.25 meters and a turning angle of 15^{o}. We report the results on the entire 'val-seen' and 'val-unseen' splits from [2].

As mentioned in Section IV-A.1, the grounding module is frozen during the training and inference. We use a Faster-RCNN model pre-trained on Visual Genome [30], to extract 20 object proposals from the RGB image observations. We use the LXMERT model adapted from huggingface [34] pre-trained on multiple multi-modal datasets (MS-COCO [35], VQA [23], GQA [36], and Visual Genome). The depth observations are separately extracted from a Resnet encoder which is updated during training. The textual instructions are tokenized to word-piece embeddings through the LXMERT tokenizer from huggingface, and then pooled out. Finally, we implement our models in PyTorch on top of AI Habitat.

VI. RESULTS

Tables I & II present a comparison of our approach against the baseline models for the Reasoning and the Vision-Language grounding experiments respectively.

A. LXMERT CMA v/s Baseline CMA

We observe that the LXMERT model marginally outperforms the baseline under the metrics of PL, NE, nDTW, SPL and ST. However, its performance drops slightly under the metrics of OS and SR. Such a moderate performance by LXMERT is counter-intuitive considering the large gains it furnishes on other Vision and Language tasks. The basic LXMERT model contains around 300 million parameters which is far more than the CMA or Seq2Seq baselines. Training LXMERT in an embodied in-simulation setting takes very long adding to the difficulty of achieving or even assessing convergence. This, limited our studies to using a pretrained LXMERT model without fine-tuning it's parameters during the VLN-CE training process. We ascribe the middling performance of LXMERT to the domain shift between the high-quality images it was pre-trained on and the significantly lower-quality visuals it experienced through the simulator. A promising future direction would be to replace the Faster RCNN from the LXMERT pipeline with a simpler, more efficient feature extractor and training the overall model on scenes from VLN-CE.

B. Seq2Seq w/ reasoning v/s Seq2Seq w/o reasoning

The agent equipped with the reasoning component achieves comparable results for the val-seen split, which contains scenes observed by the agent during training. The gains with the reasoning component are better realized for the val-unseen split, where it improves over the baseline for majority of the metrics. Although the improvements are minor, they help support our claim of the reasoning component allowing the model to generalize to unseen environments. The reasoning component described in this paper is a preliminary implementation of our idea. We plan to pursue more sophisticated mechanisms for inducing reasoning in the agent.

VII. CONCLUSIONS

In this work, we proposed the idea of a modular agent for VLN-CE. However, we focused on the high-level planning component of this agent. More specifically, we worked towards improving the baselines and presented a strategy to incorporate them into a modular architecture. Although the results are mixed and the gains are smaller, these directions appear to be promising and create ample opportunities for development in the future. Following this work, we would like to improve the Vision & Language Grounding module, by making it computationally efficient, thus allowing it to be trained or finetuned on VLN-CE episodes.

Further we would like to explore better alternatives for inculcating the ability of reasoning in such agents by allowing them to explore and understand the environments. Finally, we plan to build and test the entire modular agent by integrating the proposed high-level policy with a local planner.

ACKNOWLEDGMENT

Shaunak Halbe thanks the Bot Intelligence Group for their immense support, particularly Ingrid Navarro and Dr. Jean Oh for their mentorship and guidance. He would like to thank Felix Labelle for the invaluable inputs he offered during the project development. He is also grateful to RISS program advisors Rachel Burcin and Dr. John Dolan for their consistent support and for organizing the RISS 2021 program. Finally he would like to express his gratitude towards the sponsors for supporting him during the summer.

REFERENCES

[1] P. Anderson, Q. Wu, D. Teney, J. Bruce, M. Johnson, N. Sünderhauf, I. Reid, S. Gould, and A. van den Hengel, "Vision-and-language navigation: Interpreting visually-grounded navigation instructions in real environments," 2018.

[2] J. Krantz, E. Wijmans, A. Majumdar, D. Batra, and S. Lee, "Beyond the nav-graph: Vision-and-language navigation in continuous environments," 2020.

[3] D. Fried, R. Hu, V. Cirik, A. Rohrbach, J. Andreas, L.-P. Morency, T. Berg-Kirkpatrick, K. Saenko, D. Klein, and T. Darrell, "Speaker-follower models for vision-and-language navigation," 2018.

[4] V. Jain, G. Magalhaes, A. Ku, A. Vaswani, E. Ie, and J. Baldridge, "Stay on the path: Instruction fidelity in vision-and-language navigation," 2019.

[5] X. Wang, Q. Huang, A. Celikyilmaz, J. Gao, D. Shen, Y.-F. Wang, W. Y. Wang, and L. Zhang, "Reinforced cross-modal matching and self-supervised imitation learning for vision-language navigation," 2019.

[6] A. Majumdar, A. Shrivastava, S. Lee, P. Anderson, D. Parikh, and D. Batra, "Improving vision-and-language navigation with image-text pairs from the web," 2020.

[7] H. Tan, L. Yu, and M. Bansal, "Learning to navigate unseen environments: Back translation with environmental dropout," 2019.

[8] Y. Hong, Q. Wu, Y. Qi, C. Rodriguez-Opazo, and S. Gould, "A recurrent vision-and-language bert for navigation," 2021.

[9] A. Das, S. Datta, G. Gkioxari, S. Lee, D. Parikh, and D. Batra, "Embodied question answering," 2017.

[10] D. Gordon, A. Kembhavi, M. Rastegari, J. Redmon, D. Fox, and A. Farhadi, "Iqa: Visual question answering in interactive environments," 2018.

[11] J. Thomason, M. Murray, M. Cakmak, and L. Zettlemoyer, "Vision-and-dialog navigation," 2019.

[12] M. Shridhar, J. Thomason, D. Gordon, Y. Bisk, W. Han, R. Mottaghi, L. Zettlemoyer, and D. Fox, "Alfred: A benchmark for interpreting grounded instructions for everyday tasks," 2020.

[13] Y. Qi, Q. Wu, P. Anderson, X. Wang, W. Y. Wang, C. Shen, and A. van den Hengel, "Reverie: Remote embodied visual referring expression in real indoor environments," 2020.

[14] J. Francis, N. Kitamura, F. Labelle, X. Lu, I. Navarro, and J. Oh, "Core challenges in embodied vision-language planning," *CoRR*, vol. abs/2106.13948, 2021. [Online]. Available: https://arxiv.org/abs/2106.13948

[15] D. Misra, A. Bennett, V. Blukis, E. Niklasson, M. Shatkhin, and Y. Artzi, "Mapping instructions to actions in 3d environments with visual goal prediction," 2019.

[16] V. Blukis, D. Misra, R. A. Knepper, and Y. Artzi, "Mapping navigation instructions to continuous control actions with position-visitation prediction," 2018.

[17] H. Saha, F. Fotouhif, Q. Liu, and S. Sarkar, "A modular vision language navigation and manipulation framework for long horizon compositional tasks in indoor environment," 2021.

[18] S. Ross, G. J. Gordon, and J. A. Bagnell, "A reduction of imitation learning and structured prediction to no-regret online learning," 2011.

[19] M. Savva, A. Kadian, O. Maksymets, Y. Zhao, E. Wijmans, B. Jain, J. Straub, J. Liu, V. Koltun, J. Malik, D. Parikh, and D. Batra, "Habitat: A platform for embodied ai research," 2019.

[20] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," 2017.

[21] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," 2019.

[22] J. Lu, D. Batra, D. Parikh, and S. Lee, "Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks," 2019.

[23] A. Agrawal, J. Lu, S. Antol, M. Mitchell, C. L. Zitnick, D. Batra, and D. Parikh, "Vqa: Visual question answering," 2016.

[24] A. Das, S. Kottur, K. Gupta, A. Singh, D. Yadav, J. M. Moura, D. Parikh, and D. Batra, "Visual Dialog," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[25] H. Tan and M. Bansal, “Lxmert: Learning cross-modality encoder representations from transformers,” 2019.

[26] S. Ren, K. He, R. Girshick, and J. Sun, “Faster r-cnn: Towards real-time object detection with region proposal networks,” 2016.

[27] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” *arXiv preprint arXiv:1512.03385*, 2015.

[28] F. Xia, A. R. Zamir, Z.-Y. He, A. Sax, J. Malik, and S. Savarese, “Gibson Env: real-world perception for embodied agents,” in *Computer Vision and Pattern Recognition (CVPR), 2018 IEEE Conference on*. IEEE, 2018.

[29] A. Chang, A. Dai, T. Funkhouser, M. Halber, M. Nießner, M. Savva, S. Song, A. Zeng, and Y. Zhang, “Matterport3d: Learning from rgb-d data in indoor environments,” 2017.

[30] R. Krishna, Y. Zhu, O. Groth, J. Johnson, K. Hata, J. Kravitz, S. Chen, Y. Kalantidis, L.-J. Li, D. A. Shamma, M. S. Bernstein, and F.-F. Li, “Visual genome: Connecting language and vision using crowdsourced dense image annotations,” 2016.

[31] P. Anderson, A. Chang, D. S. Chaplot, A. Dosovitskiy, S. Gupta, V. Koltun, J. Kosecka, J. Malik, R. Mottaghi, M. Savva, and A. R. Zamir, “On evaluation of embodied navigation agents,” 2018.

[32] G. Ilharco, V. Jain, A. Ku, E. Ie, and J. Baldridge, “General evaluation for instruction conditioned navigation using dynamic time warping,” 2019.

[33] D. P. Kingma and J. Ba, “Adam: A method for stochastic optimization,” 2017.

[34] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, J. Davison, S. Shleifer, P. von Platen, C. Ma, Y. Jernite, J. Plu, C. Xu, T. L. Scao, S. Gugger, M. Drame, Q. Lhoest, and A. M. Rush, “Huggingface’s transformers: State-of-the-art natural language processing,” 2020.

[35] T.-Y. Lin, M. Maire, S. Belongie, L. Bourdev, R. Girshick, J. Hays, P. Perona, D. Ramanan, C. L. Zitnick, and P. Dollár, “Microsoft coco: Common objects in context,” 2015.

[36] D. A. Hudson and C. D. Manning, “Gqa: A new dataset for real-world visual reasoning and compositional question answering,” 2019.

[37] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov, “Proximal policy optimization algorithms,” *CoRR*, vol. abs/1707.06347, 2017. [Online]. Available: http://arxiv.org/abs/1707.06347

[38] A. Chang, A. Dai, T. Funkhouser, M. Halber, M. Niessner, M. Savva, S. Song, A. Zeng, and Y. Zhang, “Matterport3d: Learning from rgb-d data in indoor environments,” *International Conference on 3D Vision (3DV)*, 2017.

[39] E. Wijmans, A. Kadian, A. Morcos, S. Lee, I. Essa, D. Parikh, M. Savva, and D. Batra, “DD-PPO: learning near-perfect pointgoal navigators from 2.5 billion frames,” in *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net, 2020. [Online]. Available: https://openreview.net/forum?id=H1gX8C4YPr