Which way is 'right'?: Uncovering limitations of Vision-and-Language navigation models

Meera Hahn
Georgia Institute of Technology
meerahahn@gatech.edu

James M. Rehg
Georgia Institute of Technology
rehg@gatech.edu

Abstract

Vision-and-Language navigation (VLN) involves embodied agents following natural language instructions such as 'walk down the hallway and turn left at the piano.' Successful agents must be able to ground both objects referenced into the instruction such as (e.g.'piano') into the visual scene as well as ground directional phrases like (e.g.'turn left') into actions. In this work we ask the following question - to what degree are models relying on spatial and directional language cues to predict navigation? We propose a series of simple masking experiments to dissect the reliance of the models on different parts of the instruction. We surprisingly uncover that certain top performing models are only relying on the nouns of the instructions to make predictions. We additionally propose two training methods to alleviate this concerning limitation.

1 Introduction

Vision Language Navigation (VLN) is the task of having a robot navigate a visual 3D environment via following human generated natural language instructions such as *'Leave the bathroom and walk to the right'*. VLN is a popular task with multiple benchmark datasets (Anderson et al., 2018b; Ku et al., 2020; Chen et al., 2019; Qi et al., 2020; Thomason et al., 2019) which are largely conducted in simulated indoor environments such as Matterport3D (MP3D) (Chang et al., 2017) and contain thousands of human annotated instructions. The task is challenging as it requires accurate visual grounding of objects and visual descriptions provided in the instructions into the environment. Furthermore it requires the model to understand spatial language to ground instructions such as *'walk to the right'* into actions. The dataset we consider in this paper is the R2R (Anderson et al., 2018b) dataset in the discrete MP3D environments, which is constructed of a set panoramic nodes connected via navigational ability to form a navigation graph. The action space at a given time step for the navigating agent is to move to a neighboring node or to cease navigation.

Solutions to the VLN task can be divided into two distinct approaches, one which is discriminative and one which is generative. In the discriminative approach, using beam search from the given starting location, up to 30 (Majumdar et al., 2020; Guhur et al., 2021) possible paths are generated and a path is selected using a discriminative path selection model. In the generative approach the agent is placed at the starting location and the model recursively selects the next node to navigate to, until the model selects the stop action.

Both generative and discriminative approaches have seen large success in modeling by using multimodal transformer models which leverage large-scale pre-training (Majumdar et al., 2020; Guhur et al., 2021; Hong et al., 2021). The data used to pre-train these models consists of large scale web data (Sharma et al., 2018; Murahari et al., 2020; Lu et al., 2019; Hao et al., 2020) containing image-text pairs to learn visual grounding as well as large text corpora (Zhu et al., 2015) to learn linguistic semantics. Success of models for both approaches are primarily measured in terms of Success Rate (SR) which measures the percentage of selected paths that stop within 3m of the goal.

In many instances in the R2R dataset, the instructions dataset refer to the spatial layout of objects in the environment and the agents position to these objects. For example "walk past the green sofa in front of you and turn right." Due to the nature of the nodes being panoramic, learning the meaning of "in front" as well as then relating this to the "right" direction is challenging. In this paper we seek to understand to what degree the model is learning spatial and directional words and how these words impact the performance of the model. Prior work (Wang et al., 2021; Zhao et al., 2021; Zhu et al., 2021; Zhang et al., 2020; Thomason et al.,

2018) has investigated the failure modes of generative models. These works find that when making predictions, generative agents attend equally to object and direction tokens in the navigation instruction. However, there is minimal diagnostic evaluation over the discriminative VLN models.

In this paper we outline a simple method via token masking to understand how different types of part of speech and object vs direction tokens are used by discriminative models. Through our experiments we find that the discriminative models rely most heavily on nouns, almost disregarding direction tokens and other parts of speech. Additionally, we find that changing direction tokens while holding nouns tokens constant leads to no effect on the model predictions. This highlights a large limitations of these models as they are not capturing large amounts of the available information for predictions. We investigate two different training procedures in an effort to alleviate the models reliance on nouns. We hope that these findings which reveal the limitations of current VLN models will lead to new research.

Contributions:

1. We develop a diagnostic procedure to determine the influence of tokens types on VLN model's predictions.
2. We uncover the phenomena that discriminative VLN models are almost only attending to noun tokens and completely disregarding spatial tokens and all other parts of speech.
3. We propose 2 training procedures for VLN-Bert to alleviate the affects of noun token reliance as well as increase overall accuracy.

2 Masking Experiments

In this work we aim to verify that VLN models for both both the generative and discriminative approaches are learning and attending to the spatial and directional cues present in the navigational instructions. Additionally we want to quantify the degree to which these cues versus other cues are informing navigational decisions. To this end we devise a simple masking experiment over the text tokens of the instruction at inference time using the fully trained models. In each experiment we mask out tokens in accordance to their part of speech or if they fall in the category of words we define as a spatial or directional token. By testing the models performance while it doesn't have access to a specific type of token, we gain insight into the degree to which that type of token informs the model's predictions. We examine 5 different masking criterion: nouns, verbs, adjectives, left-right, spatial words. In the left-right masking experiment, only the tokens 'left' and 'right' are masked out. We add an additional experiment in which we swap all 'left' and 'right' tokens and report performance.

There is no standard list of spatial words so via qualitative analysis of the instructions among the standard VLN benchmarks we define the following words as spatial words: *'right, left, straight, near, front, through, down, up, between, past, stop, surround'*. The intuition behind these experiments is that if the model equally attends to all types of tokens performance should drop equally between different tokens being masked. Additionally we assume that instructional phrases such as 'take a left' being changed to 'take a right', should have a significant impact on performance of a navigational agent. We focus our experiments on the R2R dataset for the VLN task as it is the most widely used. Table 1 shows examples of the VLN instructions, tokenization, part of speech tagged and masked for different criterion.

We wish to examine top performing models so we pick the discriminative models: VLN-BERT (Majumdar et al., 2020) and AirBert (Guhur et al., 2021) and the generative model Recurrent-VLN-BERT (Hong et al., 2021). These models are all multi-modal transformers which use large-scale pre-training. AirBert is an extension of the VLN-BERT model and leverages and additional loss and more pre-training data scrapped from AirB&B to increase performance.

Table 1 shows the results of the masking experiments for the VLN models on the val-unseen split of the R2R dataset. Note that VLN-BERT and Airbert are discriminative models and they predicting alignment between a navigational path and the text instruction. Recurrent-VLN-BERT is a generative model in that it predicts each node in the navigation path in a iterative fashion until predicting to stop navigating.

Discriminative models rely significantly on nouns. We surprisingly observe in Table 1 that the performance of discriminative models only suffers when noun tokens are masked. Performance drops less than 2% for all other types of token masking. In fact we even see an increase in performance for VLN-BERT by up to .0034% when the 'right' and 'left' tokens are swapped to their

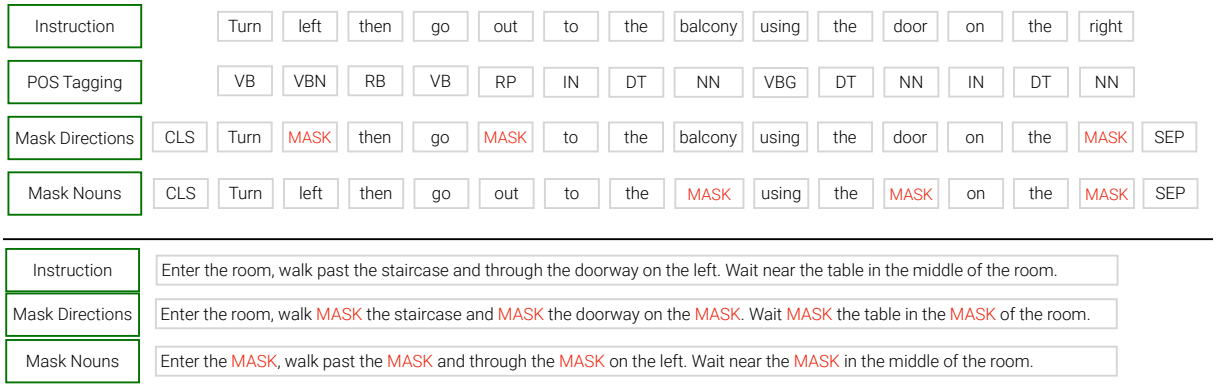

Figure 1: Example of how the navigational instructions of the R2R dataset are augmented during the masking test experiments. The instructions are part of speech tagged and tokenized. Different tokens are masked out depending on the experiment criterion. In the SWAP experiment 'left' and 'right' tokens are swapped and no token is masked.

Table 1: Results of the token type masking experiments across generative and discriminative VLN methods. The results are shown in terms of Success Rate (SR). The first column is the models performance with no augmentation to the input language.

	masking - criteria						
Model	No Masking	Swap	Left-Right	Spatial Words	Adjectives	Nouns	Verbs
VLN-BERT (Majumdar et al., 2020)	0.5926	0.5960	0.5951	0.5866	0.5922	0.4491	0.5939
AirBert (Guhur et al., 2021)	0.6645	0.6603	0.6582	0.6458	0.6582	0.4994	0.6594
Recurrent-VLN-BERT (Hong et al., 2021)	0.6275	0.4670	0.5466	0.4964	0.5917	0.4393	0.5802

antonym and when they are masked. These results indicate the models are heavily relying on noun tokens while disregarding other information. The lack of reliance on spatial words is concerning as they are an integral components of navigational language. Discriminative VLN models are presented with the entire navigation path when predicting alignment. We hypothesize that the models are relying on noun words to do pattern matching and disregarding other information and positional panoramic information of the path. We believe that the discriminative task set up for VLN may be reducing the complexity of the task such that these models can disregard most tokens in the instruction while still achieving a high SR.

Generative VLN models rely on multiple types of tokens. We find the Recurrent-VLN-BERT model suffers in performance for masking out any type of token. Masking nouns tokens still has the highest effect on performance. This could be influenced by the high density of nouns per episode in the instructions. See the Supplementary for comparison of part of speech density per episode. These results align with how we intuitively expect VLN models to perform. We also hypothesize that unlike the discriminative approach, the Recurrent-VLN-BERT is susceptible to cascading errors which will compound any drops in performance.

3 Approach

In this section we explore adding two training procedures to the VLN-BERT model with the goal of increasing the models reliance on spatial words while also increasing performance on the R2R task.

Data Augmentation We first experiment with creating meaningful hard negatives. AirBert (Guhur et al., 2021) introduced an additional shuffling training objective to effectively to teach the model to reason about temporal order. We introduce a similar shuffling strategy to the VLN-BERT training procedure. We shuffle the order of the image tokens and then use these as negative paths during the fine-tuning stage which uses the path ranking object via cross entropy loss. We report results of this model in Table 5 and refer to this in the table as shuffling.

Masked Language Modeling The VLN-BERT model is trained using masked multimodal modeling (MMM) and multimodal alignment training objectives. For MMM a randomly selected set of the input text and image tokens are masked and the model must predict the original tokens given

Table 2: Results of the token type masking experiments across different versions of the VLN-BERT model over the val-unseen split of R2R. The first column is the models performance with no augmentation to the input language.

	masking - criteria						
Model	No Masking	Swap	Left-Right	Spatial Words	Adjectives	Nouns	Verbs
VLN-BERT Original	0.5590	0.5607	0.5556	0.5564	0.5577	0.4368	0.5577
Shuffling	0.5900	0.5641	0.5824	0.5726	0.5866	0.4576	0.5896
Shuffling + SLL	0.6062	0.5743	0.6041	0.5900	0.6041	0.4334	0.5994

the surrounding context. As demonstrated by ViLBERT demonstrated that for image regions, this can be done by predicting a distribution over object classes present in the masked region. Masked text tokens however are handled the same as in BERT where the original token is predicted directly. The masked language (MLM) objective is used during stage 3 of training for VLN-BERT and masks out 15% of the tokens. However per episode there low density of spatial words and directional words such as 'left' and 'right', this is shown in the Supplementary. To encourage the model to learn the connection between directional words and the path, we train for an additional 10 epochs at the end of stage 3 where we mask out all spatial and directional words and train using only a masked language loss and do not mask out any of the image tokens. The model is then fine-tuned with the original cross-entropy loss. We report results of this training procedure in Table 5 and refer to it as the spatial language loss (SLL).

3.1 Results

Table 5 shows the results of the models trained with shuffling based negatives, the spatial language loss (SLL) and passive data in addition to the R2R train split. We find that adding shuffling increases model performance significantly, 3.1% in SR over the original model. Adding the SLL increase SR by an additional 1.62%. Note that all models have been retrained from stage 2, using ViLBERT model weights for stage 1 and 2. We found that after retraining stage 3 and 4 of the VLN-Bert model according to training specifications outlined in (Majumdar et al., 2020) we were unable to replicate the same accuracy as reported in their paper and see a 3.36% drop in SR. In Table 1 we use the VLN-Bert model provided by (Majumdar et al., 2020) which achieves higher SR.

In addition to performance over the validation sets, we seek to identify if any of these training procedures have increased the model's ability to exploit spatial information. To determine this we re-run the masking experiments over the VLN-Bert models trained with R2R, additional passive data, data augmentation and additional training objectives, results of these experiments are shown in Table 6. We find that the shuffling + SLL model shows larger reliance on spatial words than the original model. When directional words are swapped in the original model, performance increases slightly. In contrast the shuffling + SLL model suffers a drop in performance of 3.19%, which is the largest performance drop for the masking experiments outside of noun masking. Based on the results of these masking experiments we can assume the shuffling + SLL model is leveraging more spatial and directional information than the original model.

Table 3: Results of the VLN-Bert architecture trained with different data augmentation and training objectives. Note that all models have been retrained from stage 2, using ViLBERT model weights for stage 1 and 2.

	val-seen			val-unseen		
Method	NE ↓	SR ↑	SPL ↑	NE ↓	SR ↑	SPL ↑
Original Loss	**4.1093**	**0.6667**	**0.6309**	4.6699	0.5590	0.5158
Shuffling	4.2836	0.6608	0.6209	4.5222	0.5900	0.5436
Shuffling + SLL	4.3506	0.6490	0.6105	**4.1503**	**0.6062**	**0.5608**

4 Conclusion

In this work we highlight a significant limitation of the discriminative VLN models. We propose a set of token masking experiments over VLN models to infer what parts of the text instructions the models attend to. Via these experiments we find that the most popular discriminative VLN models seem to solely rely on the nouns of the navigation instructions. In order to encourage the discriminative models to leverage other parts of the instruction we try using an additional training objective and data augmentation and find these strategies to be effective and lead to higher performance on the VLN task.

5 Supplementary

5.1 Comparison of Dataset Language

In Table 4 we compare the language of instructions between existing VLN datasets as to understand the composition of token types in the navigational instructions. Specifically we compare: dataset size, vocab size, average text length per episode, and density of different parts of speech (POS) and spatial tokens per episode. Vocab size was determined by the total number of unique words. We used the (Loper and Bird, 2002) POS tagger to calculate the POS densities over the text in each dataset. We note that the RxR task has significantly longer instructions than any of the other datasets. When looking at the density of different parts of speech, we find that all datasets have a high density of nouns and lower density of adjectives and verbs. There is also a low density of spatial words like "left" and "right", which is to be expected as they are maybe only used a few times in an instruction.

5.2 Counterfactuals

One reason that performance would not be effected when swapping 'left' and 'right' tokens or masking them out, could be that many paths simply don't have a counterfactual. In other words, when an instruction states to 'turn right' it is possible there is no option to turn left in the discrete panoramic node setting of MP3D (Chang et al., 2017) of the R2R dataset. For example, imagine the navigation agent exits a room to find a hallway where only neighboring nodes are to the right of the agent. To discredit this hypothesis we first look at each turn in the R2R paths for the val-unseen data split. We determine any turn to be when the heading of the agent changes by over 30°degrees between two nodes in the path. We find there are an average of 1.67 turns per episode. Then for each node where the agent makes a turn we determine if there are any neighboring nodes which the agent could have navigated to instead, which turn in the opposite way or go straight. We call these nodes counterfactuals and we find that per turn there is an average of 1.62 counterfactuals. This discredits the possibility that directional tokens do not serve a significant role in the VLN task.

Table 4: Comparison of the language between the common VLN benchmark datasets: Reverie (Qi et al., 2020), RXR (Ku et al., 2020) and R2R (Anderson et al., 2018b). Compares the size of the datasets and density of different POS.

	Reverie	RXR	R2R
Dataset Size	21702	25368	21582
Vocab Size	4815	3779	3999
Avg Num Tokens	18.3073	97.2956	29.3665
Noun Density	0.3155	0.2104	0.2775
Adj Density	0.0505	0.0615	0.0461
Verb Density	0.1163	0.1690	0.1222
Left-Right Density	0.0487	0.0492	0.0664

5.3 Training via Passive Data

Discriminative transformer based models for VLN rely heavily on noun tokens while disregarding other types of tokens. Directional and spatial tokens provide significant information that these models are currently not taking advantage of. To combat model reliance on tokens that describe object and rooms, we seek to inject new data during the training stage on R2R which contains minimal references to object and room names. We generate the new paths and the instructions programmatically allowing us to forgo the need for additional human annotations.

To generate the additional navigational paths we use a similar strategy to that of (Anderson et al., 2018a). We first sampled start and goal location pairs in the MP3D training environments and then found the shortest path on the scene's navigation graph. We discarded any paths that were contained in the R2R dataset and any paths with less than 5 edges and over 10 edges. We then generated instructions over the paths. If a path had a turn of over 120°degrees we generated the instruction "turn around" otherwise any turn over 30°degrees generated the instruction "turn (left | right)". Otherwise instructions such as "go straight", "go forward", "continue straight", etc for "x meters" were generated, where x was determined to meters for the straight region of the path. If the path navigated over stairs we generated an instruction to "go (up|down) the stairs". Each generated instruction ended with a "stop" or "wait here" command. In total we generate an additional 6k instruction-path pairs over the training environments of MP3D. For example a generated instruction in our dataset is *'Go forward and walk 3 meters. Turn right, and walk one meter. Stop.'*

We then add the new generated path-instruction pairs to the training split of the R2R dataset. We retrain the VLN-BERT model with the additional data for stage 3 and then fine-tune only on the origi-

nal RxR train split. In stage 3 of training the model is being trained by the masked multi-modal modeling objectives. The results of the model trained with additional passive data is shown in Table 5. We find that adding passive data to the training procedure does not significantly change performance over the R2R dataset validation sets for the VLN task. We then run the masking experiments to determine if adding passive data to the training data helps increase spatial token reliance. We find that passive data does not help and the masking experiments follow that of the original loss and data.

Table 5: Results of the VLN-Bert architecture trained with different data augmentation and training objectives. Note that all models have been retrained from stage 2, using ViLBERT model weights for stage 1 and 2.

	val-seen			val-unseen		
Method	NE ↓	SR ↑	SPL ↑	NE ↓	SR ↑	SPL ↑
Original Loss	**4.1093**	**0.6667**	**0.6309**	4.6699	0.5590	0.5158
Shuffling	4.2836	0.6608	0.6209	4.5222	0.5900	0.5436
Shuffling + SLL	4.3506	0.6490	0.6105	**4.1503**	**0.6062**	**0.5608**
R2R + Passive Data	**3.8264**	**0.6804**	**0.6390**	4.7218	0.5453	0.5000

References

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

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

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

Howard Chen, Alane Suhr, Dipendra Misra, Noah Snavely, and Yoav Artzi. 2019. Touchdown: Natural language navigation and spatial reasoning in visual street environments. In *CVPR*.

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

Weituo Hao, Chunyuan Li, Xiujun Li, Lawrence Carin, and Jianfeng Gao. 2020. Towards learning a generic agent for vision-and-language navigation via pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13137–13146.

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

Alexander Ku, Peter Anderson, Roma Patel, Eugene Ie, and Jason Baldridge. 2020. Room-across-room: Multilingual vision-and-language navigation with dense spatiotemporal grounding. *arXiv preprint arXiv:2010.07954*.

Edward Loper and Steven Bird. 2002. Nltk: the natural language toolkit. *arXiv preprint arXiv:0205028*.

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

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

Vishvak Murahari, Dhruv Batra, Devi Parikh, and Abhishek Das. 2020. Large-scale pretraining for visual dialog: A simple state-of-the-art baseline. In *European Conference on Computer Vision*. Springer.

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

Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. 2018. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*.

Jesse Thomason, Daniel Gordon, and Yonatan Bisk. 2018. Shifting the baseline: Single modality performance on visual navigation & qa. *arXiv preprint arXiv:1811.00613*.

Jesse Thomason, Michael Murray, Maya Cakmak, and Luke Zettlemoyer. 2019. Vision-and-dialog navigation. In *CoRL*.

Su Wang, Ceslee Montgomery, Jordi Orbay, Vighnesh Birodkar, Aleksandra Faust, Izzeddin Gur, Natasha Jaques, Austin Waters, Jason Baldridge, and Peter Anderson. 2021. Less is more: Generating grounded navigation instructions from landmarks. *arXiv preprint arXiv:2111.12872*.

Table 6: Results of the token type masking experiments across different versions of the VLN-BERT model over the val-unseen split of R2R. The first column is the models performance with no augmentation to the input language.

	masking - criteria						
Model	No Masking	Swap	Left-Right	Spatial Words	Adjectives	Nouns	Verbs
VLN-BERT Original	0.5590	0.5607	0.5556	0.5564	0.5577	0.4368	0.5577
Shuffling	0.5900	0.5641	0.5824	0.5726	0.5866	0.4576	0.5896
Shuffling + SLL	0.6062	0.5743	0.6041	0.5900	0.6041	0.4334	0.5994
Add Passive Data	0.5453	0.5398	0.5445	0.5407	0.5364	0.4027	0.5364

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

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

Wanrong Zhu, Yuankai Qi, Pradyumna Narayana, Kazoo Sone, Sugato Basu, Xin Eric Wang, Qi Wu, Miguel Eckstein, and William Yang Wang. 2021. Diagnosing vision-and-language navigation: What really matters. *arXiv preprint arXiv:2103.16561*.

Yukun Zhu, Ryan Kiros, Rich Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In *Proceedings of the IEEE international conference on computer vision*.