Diagnosing Vision-and-Language Navigation: What Really Matters

Wanrong Zhu[¶], Yuankai Qi[§], Pradyumna Narayana[*], Kazoo Sone[*], Sugato Basu[*], Eric Xin Wang[‡], Qi Wu[§], Miguel Eckstein[¶], William Yang Wang[¶]
[¶]UC Santa Barbara, [§]University of Adelaide, [*]Google, [‡]UC Santa Cruz
{wanrongzhu,miguleckstein,wangwilliamyang}@ucsb.edu, qykshr@gmail.com
{pradyn,sone,sugato}@google.com, xwang366@ucsc.edu, qi.wu01@adelaide.edu.au

Abstract

Vision-and-language navigation (VLN) is a multimodal task where an agent follows natural language instructions and navigates in visual environments. Multiple setups have been proposed, and researchers apply new model architectures or training techniques to boost navigation performance. However, there still exist non-negligible gaps between machines' performance and human benchmarks. Moreover, the agents' inner mechanisms for navigation decisions remain unclear. To the best of our knowledge, how the agents perceive the multimodal input is under-studied and needs investigation. In this work, we conduct a series of diagnostic experiments to unveil agents' focus during navigation. Results show that indoor navigation agents refer to both object and direction tokens when making decisions. In contrast, outdoor navigation agents heavily rely on direction tokens and poorly understand the object tokens. Transformer-based agents acquire a better cross-modal understanding of objects and display strong numerical reasoning ability than non-Transformer-based agents. When it comes to vision-and-language alignments, many models claim that they can align object tokens with specific visual targets. We find unbalanced attention on the vision and text input and doubt the reliability of such cross-modal alignments.[1]

1 Introduction

A key challenge for Artificial Intelligence (AI) research is to move beyond Independent and Identically Distributed (i.i.d.) data analysis: We need to teach AI agents to understand multimodal input data, and jointly learn to reason and perform incremental and dynamic decision-making with the help from humans. Vision-and-Language Navigation (VLN) has received much attention due to its active perception and multimodal grounding setting, dynamic decision-making nature, rich applications, and accurate evaluation of agents' performances in language-guided visual grounding. As the AI research community gradually shifts its attention from the static empirical analysis of datasets to more challenging settings that require incremental decision-making processes, the interactive task of VLN deserves a more in-depth analysis of why it works and how it works.

	R2R	RxR	Touchdown
Human Performance	86	94	92
SoTA Model Performance	78	53	17

Table 1: There exists salient gaps between machines' vision-and-language navigation (VLN) performance and human benchmarks. Navigation success rates are reported on the R2R (Anderson et al., 2018) and the RxR dataset (Ku et al., 2020b) for indoor VLN and the Touchdown dataset (Chen et al., 2019) for outdoor VLN.[2]

Various setups have been proposed to address to the VLN task. Researchers generate visual trajectories and collect human-annotated instructions for indoor (Anderson et al., 2018; Jain et al., 2019a; Ku et al., 2020a; Chen et al., 2021) and outdoor environment (Chen et al., 2019; Mehta et al., 2020; Mirowski et al., 2018). There are also interactive VLN settings based on dialogues (Nguyen et al., 2019; Nguyen and III, 2019; Zhu et al., 2020c), and task that navigates agents to localize a remote object (Qi et al., 2020c). However, few studies ask the *Why* and *How* questions: Why do these agents work (or do not work)? How do agents make decisions in different setups?

Through the years, agents with different model architectures and training mechanisms have been proposed for indoor VLN (Anderson et al., 2018; Fried et al., 2018; Hao et al., 2020; Hong et al., 2020a,b; Huang et al., 2019; Ke et al., 2019; Li et al., 2019; Ma et al., 2019a; Qi et al., 2020b; Tan

[1]Code and data used in this study are available at https://github.com/VegB/Diagnose_VLN.

[2]We record the published state-of-the-art performance on R2R, RxR and Touchdown leaderboards on Dec.15th, 2021.

Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 5981 - 5993
July 10-15, 2022 ©2022 Association for Computational Linguistics

et al., 2019; Wang et al., 2020a, 2019, 2018, 2020b; Zhu et al., 2020a) and outdoor VLN (Chen et al., 2019; Ma et al., 2019b; Mirowski et al., 2018; Xia et al., 2020; Xiang et al., 2020; Zhu et al., 2020b). Back-translation eases the urgent problem of data scarcity (Fried et al., 2018). Imitation learning and reinforcement learning enhance agents' generalization ability (Wang et al., 2019, 2018). With the rise of BERT-based models, researchers also apply Transformer and pre-training to further improve navigation performance (Hao et al., 2020; Hong et al., 2020b; Zhu et al., 2020b). While applying new techniques to the navigation agents might boost their performance, we still know little about how agents make each turning decision. Treatment of the agents' processing of instructions and perception of the visual environment as a black box might hinder the design of a generic model that fully understands visual and textual input regardless of VLN setups. Table 1 shows non-negligible performance gaps between neural agents and humans on both indoor and outdoor VLN tasks.

Therefore, we focus on analyzing how the navigation agents understand the multimodal input data in this work. We conduct our investigation from the perspectives of natural language instruction, visual environment, and the interpretation of vision-language alignment. We create counterfactual interventions to alter the instructions and the visual environment in the validation dataset, focusing on variables related to objects, directions and numerics. More specifically, we modify the instruction by removing or replacing the object/direction/numeric tokens, and we adjust the environment by masking out visual instances or horizontally flipping the viewpoint images. Subsequently, we examine the interventions' treatment effects on agents' evaluation performance while keeping other variables unchanged. We set up experiments on the R2R (Anderson et al., 2018) and the RxR dataset (Ku et al., 2020b) for indoor VLN and the Touchdown dataset (Chen et al., 2019) for outdoor VLN. We examine nine VLN agents on the three datasets with quantitative ablation diagnostics on the text and visual inputs.

In summary, our key findings include:

1. Indoor navigation agents refer to both objects and directions in the instruction when making decisions. In contrast, outdoor navigation agents heavily rely on directions and poorly understand visual objects. (Section 4)
2. Instead of merely staring at surrounding objects, indoor navigation agents are able to set their sights on objects further from the current viewpoint. (Section 5)
3. Transformer-based agents display stronger numerical reasoning ability (Section 4), and acquire better cross-modal understanding of objects, compared to non-Transformer-based agents. (Section 6)
4. Indoor agents can align object tokens to certain targets in the visual environment to a certain extent, but display in-balanced attention on text and visual input. (Section 6)

We hope these findings reveal opportunities and obstacles of current VLN models and lead to new research directions.

2 Related Work

Instruction Following is a long-standing topic in AI studies that ask an agent to follow natural language instructions and accomplish target tasks, which can be dated back to the SHRLDU (Winograd, 1971). Efforts made to tackle this classic problem spans from defining templates (Klingspor et al., 1997; Antoniol et al., 2011), designing hard-encoded concepts to ground visual attributes and spatial relations (Steels and Vogt, 1997; Roy, 2002; Guadarrama et al., 2013; Kollar et al., 2013; Matuszek et al., 2014), to constructing varies datasets and learning environments (Anderson et al., 1991; Bisk et al., 2016a; Misra et al., 2018). Many methods have been proposed to map the instructions into sequence of actions, such as reinforcement learning (Branavan et al., 2009, 2010; Vogel and Jurafsky, 2010; Misra et al., 2017), semantic parsing (Chen and Mooney, 2011; Artzi and Zettlemoyer, 2013), alignment-based model (Andreas and Klein, 2015), and neural networks (Bisk et al., 2016b; Mei et al., 2016; Tan and Bansal, 2018).

Vision-and-Language Navigation is a task where an agent comprehends the natural language instructions and reasons through the visual environment. Many studies aim at improving VLN agents' performance in one way or another. To enrich training data, a line of work (Fried et al., 2018; Zhu et al., 2020b) use back-translation to generate augmented instructions. To enforce cross-modal grounding, RPA and RCM (Wang et al., 2018, 2019) use reinforcement learning, SMNA (Ma et al., 2019a) uses a visual-textual co-grounding module to improve cross-modal alignment, Rel-

Graph (Hong et al., 2020a) uses graphs for task formulation. To address the generalizability problem to unseen environment, PRESS (Li et al., 2019) introduces a stochastic sampling scheme, EnvDrop (Tan et al., 2019) proposes environment dropout. To utilize visual information from the environment, AuxRN (Zhu et al., 2020a) uses auxiliary tasks to assist semantic information extraction, VLN-HAMT (Chen et al., 2021) incorporates panorama history with a hierarchical vision transformer. FAST (Ke et al., 2019) makes use of asynchronous search and allows the agent to backtrack if it discerns a mistake after attending to global and local knowledge. With the success of BERT-related models in NLP, researchers also start to build Transformer-based navigation agents and add a pre-training process before fine-tuning on the downstream VLN task (Hao et al., 2020; Hong et al., 2020b; Zhu et al., 2020b; Chen et al., 2021).

Model Behavior Analysis As multimodal studies gain more and more attention, there are lines of works that focus on explaining models' behaviors to better understand and handle the tasks. Some generate textual explanations by training another model to mimic human explanations (Hendricks et al., 2016; Park et al., 2018; Wu and Mooney, 2019). Others generate visual explanations with the help of attention mechanism (Lu et al., 2016) or gradient analysis (Selvaraju et al., 2017). There are also attempts to provide multimodal explanations, e.g., Li et al. (2018) breaks up the end-to-end VQA process and examines the intermediate results by extracting attributes from the visual instances. Another line of work examines model performance by conducting ablation studies on input data. Recent analyses on language modelling (O'Connor and Andreas, 2021), machine translation (Fernandes et al., 2021), and instruction following (Dan et al., 2021) ablate/perturb both training and validation data. A study on multimodal models (Frank et al., 2021) only applies ablation during evaluation, which is the same as our settings.

3 Background and Research Questions

We first bring in the task of Vision-and-Language Navigation and introduce the datasets and agents used for comparison. Then we list out the research questions to study in this work.

Dataset	Model	Trans?	Visual Feature
R2R	EnvDrop (Tan et al., 2019)	×	ResNet-152
	FAST (Ke et al., 2019)	×	
	VLN ⟳ BERT (Hong et al., 2020b)	✓	
	PREVALENT (Hao et al., 2020)	✓	
RxR-en	CLIP-ViL (Shen et al., 2021)	×	CLIP-ViT
	VLN-HAMT (Chen et al., 2021)	✓	
Touchdown	RCONCAT (Chen et al., 2019)	×	ResNet-18
	ARC (Xiang et al., 2020)	×	
	VLN-Transfomer (Zhu et al., 2020b)	✓	

Table 2: The VLN datasets and models covered in this study. We record whether the model structure is Transformer-based, and the pre-trained feature extractor used to encode visual environment.

3.1 Vision-and-Language Navigation

In the vision-and-language navigation task, the navigation agent is asked to find the path to reach the target location following the instructions $\mathcal{X}$. The navigation procedure can be viewed as a sequential decision-making process. At each time step t, the visual environment presents an image view $\boldsymbol{v}_t$. With reference to the instruction $\mathcal{X}$ and the visual view $\boldsymbol{v}_t$, the agent is expected to choose an action a_t such as *turn left* or *stop*.

Datasets We conduct indoor navigation experiments on the Room-to-Room (R2R) dataset (Anderson et al., 2018) and the Room-across-Room (RxR) dataset (Ku et al., 2020b), and test outdoor VLN on Touchdown (Chen et al., 2019). R2R and RxR are built upon real estate layouts and contain separate graphs for each apartment/house. Unlike R2R, which shoots for the shortest path, RxR has longer and more variable paths. R2R only contains English instructions, while RxR also includes instructions in Hindi and Telugu. In this study, we only cover the English subset for RxR, and will refer to it as RxR-en in the following sections. Navigation in Touchdown occurs in the urban environment, where the viewpoints form a huge connected graph. Compared to indoor environments, Touchdown has more complicated visual environments and a more extensive search space. The evaluation results are reported on the validation unseen sets for R2R and RxR-en and on the test set for Touchdown.

Models Table 2 lists out the models covered in our study. We use the code and trained checkpoints shared by the authors in the following experiments.

For indoor navigation on R2R, we study a widely adopted base model Envdrop (Tan et al., 2019), a backtracking framework for self-correction FAST (Ke et al., 2019), and two SoTA models VLN ⟳ BERT (Hong et al., 2020b) and PREVA-

LENT (Hao et al., 2020). The Envdrop introduces environment dropout on top of the Speaker-Follower (Fried et al., 2018) model, FAST conducts an asynchronous search for backtracking, PREVALENT, and VLN ⟳ BERT are Transformer-based agents with pre-trained models.

For navigation on RxR-en, we examine CLIP-ViL (Shen et al., 2021) and VLN-HAMT (Chen et al., 2021). CLIP-ViL shares the same model structure with EnvDrop. The only difference is that CLIP-ViL uses CLIP-ViT (Radford et al., 2021) to extract visual features, while EnvDrop uses ImageNet ResNet (Szegedy et al., 2017) features. VLN-HAMT incorporates a long-horizon history into decision-making by encoding all the past panoramic observations via a hierarchical vision Transformer.

For outdoor navigation on Touchdown, we consider the common baseline RCONCAT (Chen et al., 2019), and two SoTA models ARC (Xiang et al., 2020) and VLN-Transfomer (Zhu et al., 2020b). RCONCAT encodes the trajectory and the instruction in an LSTM-based manner. ARC improves RCONCAT by paying special attention to the stop signals. VLN-Transfomer is a Transformer-based agent that applies pre-training on an external dataset for outdoor navigation in urban areas.

Metrics In the following experiments, we evaluate navigation performance with Success Rate (SR) for indoor agents and Task Completion (TC) rate for outdoor agents. Both SR and TC measure the accuracy of completing the navigation task, reflecting the agents' overall ability to finish navigation correctly. An indoor navigation task is considered complete if the agent's final position locates within 3 meters of the target location. For outdoor navigation, the task is considered complete if the agent stops at the target location or one of its adjacent nodes in the environment graph.

3.2 Research Questions

Current VLN studies have reached their bottleneck as only minor performance improvements have been achieved recently, while a significant gap still exists between machine and human performance. This motivates us to find the reasons.

To better understand how VLN agents make decisions during navigation, we conduct a series of experiments on indoor and outdoor VLN tasks, aiming to answer the following questions that might help us locate the deficiencies of current model designs and explore future research directions:

1. *What can the agents learn from the instructions? Do they pay more attention to object tokens or directions tokens? Do they have the ability to count? (Section 4)*
2. *What do agents see in the visual environment? Are they staring at the closely surrounded objects or also browsing further layout? Do they focus on individual visual instances or perceive the overall outline? (Section 5)*
3. *Can agents match textual tokens to visual entities? How reliable are such connections? (Section 6)*

Dataset	Instruction
R2R	Walk through the **door** by the **sink** into the **middle** of the next **room**. Turn ***right*** and walk down the **hallway** and enter the **third door** on your ***right***.
RxR-en	We're facing towards a small **picture** that's attached to the **wall**, turn slightly to the ***right***, and enter the **hallway** that's in ***front*** of you, turn to the ***left***, take **five steps** further... On your ***right*** there are a few **glass doors** and on the ***left*** there's a **living room**, walk towards the **living room**, turn slightly to the ***left***... On the ***right*** there are **four chairs** and a beautiful **coffee table** in the **middle**, on the ***left*** there's a **console table** with a **vase** with **flowers** on **top**, walk past the **console table** towards the ***back*** of the **chair** that's in direcfront of you... We're now facing towards a **lamp**, and on the ***right*** there's a marble **console table** with **decorations** on **top**, and that's your **destination**.
Touchdown	Orient yourself so that you are moving in the same **direction** as **traffic**. Go straight through **3 intersections**. Keep moving ***forward***, after the **3rd intersection**, you should see a **signs** for a **store** with a white **background** and red **dots** as well as a red and white **bullseye target**. Continue going straight past this **store** and at the next **intersection**, turn ***left***. Go through **one intersection** and ***stop*** just after the **wall** on your ***left*** with the purple **zig zag patterns**.

Table 3: Instructions from R2R, RxR-en and Touchdown with **object-tokens**, ***direction-tokens*** and **numeric-tokens** highlighed.

Dataset	#Data	$\overline{L_{\text{path}}}$	$\overline{L_{\text{instr}}}$	#Object	$\overline{p(\text{tok}_{\text{obj}})}$	$\overline{p(\text{tok}_{\text{direc}})}$
R2R	2.3k	6.0	29.3	0.6k	19.8%	7.3%
RxR-en	4.6k	8.5	111.3	1.4k	16.1%	6.5%
Touchdown	1.4k	34.4	92.5	1.0k	16.8%	6.8%

Table 4: Statistics of R2R, RxR-en and Touchdown datasets. #Data is the dataset size used for evaluation in this study. $\overline{L_{\text{path}}}$ is the average path length, which is the number of viewpoints covered in the trajectory. $\overline{L_{\text{instr}}}$ is the average instruction length. #Object denotes the number of unique objects mentioned in the instructions. $\overline{p(\text{tok}_{\text{obj}})}$ and $\overline{p(\text{tok}_{\text{direc}})}$ indicates the percentage of object/direction tokens per instruction.

4 Analysis on Instruction Understanding

This section examines whether and to what extent the agent understands VLN instructions. We focus on how the agent perceives object-related tokens, direction-related tokens, and numeric tokens, and their effects on final navigation performance. Table 3 shows exemplar instructions of the three

*	Setting	Instruction
1	Vanilla	Go ***left*** down the **hallway** toward the **exit sign**. Go into the **door** on the ***left*** and ***stop*** by the **table**.
2	Mask Object Tokens	Go left down the **[MASK]** toward the **[MASK] [MASK]**. To into the **[MASK]** on the left and stop by the **[MASK]**.
3	Replace Object Tokens	Go left down the **portrait** toward the **sofa fountains**. Go into the **football** on the left and stop by the **boats**.
4	Controlled Trial	Go **[MASK]** down the hallway **[MASK]** the exit sign. To into the door on **[MASK]** left and **[MASK]** by **[MASK]** table.
5	Mask Direction Tokens	Go ***[MASK]*** down the hallway toward the exit sign. Go into the door on the ***[MASK]*** and ***[MASK]*** by the table.
6	Replace Direction Tokens	Go ***right*** down the hallway toward the exit sign. Go into the door on the ***right*** and ***forward*** by the table.
7	Controlled Trial	Go left down the ***[MASK] [MASK]*** the exit sign. Go into the door on the left and ***[MASK]*** by the table.

Table 5: Example of instruction modification. In the original instruction, there are five **object-related tokens**, and three ***direction-related tokens***. In the object token ablations, we mask out the object tokens, or replace them with randomly sampled object tokens. The controlled trial randomly masked out five tokens from the instruction for a fair comparison. Likewise the direction tokens.

*	Ablation	Setting	SR ↑ on R2R				SR ↑ on RxR-en		TC ↑ on Touchdown		
			EnvDrop	FAST	VLN⟳BERT	PREVALENT	CLIPViL	HAMT	RCONCAT	ARC	VLNTrans
1	–	Vanilla	49.77	63.90	53.30	57.13	40.21	52.52	11.78	15.19	16.11
2	Object	Mask	-36%	-38%	-21%	-20%	-48%	-32%	-34%	-36%	-6%
3		Controlled Trial	-31%	-26%	-8%	-8%	-35%	-23%	-44%	-55%	-15%
4	Direction	Mask	-23%	-23%	-15%	-11%	-39%	-28%	-73%	-90%	-45%
5		Controlled Trial	-11%	-12%	-4%	-3%	-14%	-9%	-22%	-23%	-8%

Table 6: The navigation performance for indoor and outdoor agents on object-token and direction-token ablations. We record the validation score in the "vanilla" setting, and report the relative performance change for each ablation setting. For object-token ablations, the "mask" setting masks out all the object-tokens, while the controlled trial masks out the same amount of random tokens. The same applies to direction-token ablations.

datasets covered in our study. As shown in Table 4, Touchdown's trajectory length is significantly longer than the other two indoor datasets. RxR-en and Touchdown have longer instructions than R2R. The ratios of object and direction tokens in all three datasets are comparable, involving about two times more object tokens than direction tokens.

4.1 The Effect of Object-related Tokens

We first create counterfactual interventions on instructions by masking out the object tokens. We use Stanza (Qi et al., 2020a) part-of-speech (POS) tagger to locate object-related tokens. A token will be regarded as an object token if its POS tag is *NOUN* or *PROPN*. During masking, we replace the object token with a specified mask token *[MASK]*. Then we examine the average treatment effects of the intervention on agents' performance, while keeping other variables unchanged.

Noticeably, when we mask out the object tokens, the number of visible tokens in the provided instruction also decreases, which is a coherent factor with masking object tokens and might interfere with our analysis. To eliminate the effect of reducing visible tokens, we add a controlled trial in which we randomly mask out the same amount of tokens. Table 5 gives an example of masking object tokens (#2) and its corresponding controlled trial (#4).

We follow each agent's original experiment setting for all the experiments in this study and train it on the original train set. Then we apply masking to object tokens in the validation set, and report agents' relative performance changes under each setting. We conduct five repetitive experiments and report the average scores for settings that involve random masking or replacing.

Table 6 presents how the agents' navigation performance change when object tokens are masked out (#2 & #3). Intuitively, not knowing what objects are mentioned in the instruction lowers all models' performance. Comparing the masking ablations with the controlled trial for indoor VLN, we notice that masking out the object tokens result in a more drastic decrease in success rate than masking out random tokens. This holds for all indoor agents, which verifies that indoor agents depend on object tokens more than other tokens. However, when we compare results on the Touchdown for outdoor VLN, we notice in surprise that masking out the object tokens has a weaker impact on task completion rate than masking out random tokens. This suggests that current outdoor navigation agents do not fully take object tokens into consideration when making decisions. This may be caused by the weak visual recognition module in current outdoor agents. As addressed in Table 2, all three outdoor

agents rely on visual features extracted by ResNet-18, which may not be powerful enough to fully incorporate the complicated urban environments.

4.2 The Effect of Direction-related Tokens

We regard the following tokens as direction-related tokens: *left, right, back, front, forward, stop*. Similar to how we ablate the object tokens, we mask out direction tokens from the instruction and examine the impact on agents' navigation performance. Table 5 provides examples of direction tokens masking (#5), and its controlled trial (#7) where the same amount of random tokens are masked out. Table 6 shows agents' performance under various direction tokens ablation settings (#4 & #5).

For indoor agents, masking out the direction tokens cause a sharper drop in success rate compared to masking out random tokens, which means the indoor navigation agents do consider the direction tokens during navigation. We also notice that agents are more sensitive to the loss of direction guidance on RxR-en than on R2R. Such difference may be caused by the way these two datasets are designed. R2R's ground-truth trajectories are the shortest path from start to goal. Previous studies have noted that R2R has the danger of exposing structural bias and leaking hidden shortcuts (Thomason et al., 2019), and that such design encourages goal-seeking over path adherence (Jain et al., 2019b). RxR is crafted to include longer and more variable paths to avoid such biases. Naturally, agents on RxR-en would pay more attention to direction tokens since they may approach their goal indirectly.

For outdoor navigation agents, masking out direction tokens leads to a drastic decline in task completion rate, compared to random masking. This indicates that current outdoor navigation agents heavily rely on the direction tokens when making decisions. Given the complicated visual environments and instructions in the outdoor navigation task, current agents fail to fully use the instructions, especially ignoring the rich object-related information. The ARC model shows the most salient performance decline of 90% to the instructions ablated by direction token masking. Aside from the classifier that predicts the next direction to take, ARC also uses a stop indicator to decide whether to stop at each step or not. Its unique mechanism for detecting stop signals might explain why it is more sensitive to the existence of direction tokens.

	#Data	$\overline{\text{L}_{\text{instr}}}$	$\overline{p(\text{tok}_{\text{num}})}$
RxR-en	2.0k	135.0	1.4%
Touchdown	1.0k	100.1	2.0%

Table 7: Statistics of RxR-en and Touchdown data samples with numeric tokens examined for evaluation. $\overline{\text{L}_{\text{instr}}}$ is the average instruction length, and $\overline{p(\text{tok}_{\text{num}})}$ denotes the percentage of numeric tokens per instruction.

Setting	SR ↑ on RxR-en		TC ↑ on Touchdown		
	CLIPViL	HAMT	RCONCAT	ARC	VLNTrans
Vanilla	36.05	47.38	11.76	14.24	16.31
Mask Number	-4%	-5%	3%	-7%	-3%
Replace Number	-3%	-6%	1%	2%	-11%
Controlled Trial	-6%	-3%	-5%	-4%	-6%

Table 8: Navigation performance on different numeric-token ablations settings.

4.3 The Effect of Numeric Tokens

We conduct ablation studies on agents' understanding of numeric tokens on RxR-en for indoor agents and Touchdown for outdoor agents. We select a subset of examples whose instructions contain numeric tokens,[3] and construct ablated instructions on top. Table 7 provides the statistics of the instructions for numeric ablations. Table 8 lists out the results. The VLN-HAMT on RxR-en and VLN-Transformer on Touchdown have comparable performance when masking numeric tokens over random tokens, and have worse performance when replacing numeric tokens. This suggests that these two Transformer-based agents have the ability to conduct numerical reasoning to some extent. In contrast, other non-Transformer-based agents have less salient performance drops when replacing numeric tokens. For RCONCAT and ARC, replacing numeric tokens even leads to higher task completion rates. This implies the insufficient counting ability of the non-Transformer-based agents.

5 Analysis on Visual Environment

This section investigates what the agent perceives in the visual environment. We set an eye on inspecting the agent's understanding of the surrounding objects and direction-related information.

5.1 Effect of Objects in the Environment

Built upon the Matterport dataset (Chang et al., 2017), R2R and RxR obtain detailed object instance annotations and serve as an excellent source

[3] We consider instructions that contain cardinal numbers from 1 to 20, and ordinal numbers from 1st to 20th.

(a) Foreground Objects (b) All Visible Objects

Figure 1: Accessible objects within different ranges.

Setting	R2R	RxR-en
All Visible Objects (except wall/floor/ceiling)	35.1	35.5
Foreground Objects	3.2	3.9
Objects Mentioned in Instruction	5.3	11.8

Table 9: The average number of visual objects in the panorama at each viewpoint under different settings.

for our visual object studies. Touchdown is based on Google Street View and does not acquire object-related annotations. Thus, we conduct experiments on the indoor VLN environment.

We designed several ablation settings for visual objects. The "mask all visible" setting applies masking to all the visible visual objects in the environment (except for wall/ceiling/floor). The "mask foreground" setting ablates the visual objects within 3 meters of the camera viewpoint, which we refer to as the foreground area. The region beyond 3 meters from the camera viewpoint is regarded as the background area. Figure 1 shows an example for comparison. We choose 3 meters as the boundary because the bounding box annotations for objects within 3 meters are provided in REVERIE (Qi et al., 2020c). We denote the number of visual objects within 3 meters as k, and add a controlled trial that masks out k random visual objects from all the visible objects at the current viewpoint, regardless of their depth.

Table 9 compares the number of visual objects under various ablation settings. We mask out the objects in each view by filling the corresponding bounding boxes with the mean color of the surrounding. Then we follow original experiment settings and use ResNet-152 (He et al., 2016) CNN to extract image features for R2R agents, and use CLIP-ViT-B/32 (Radford et al., 2021) to extract visual features for RxR-en agents.

Results for visual object ablations are shown in Table 10. We examine the influence of masking out different quantities of visual objects by comparing the "mask all visible" setting with the controlled trial (#2 vs. #4). It comes naturally that masking

*	Ablation	Setting	SR ↑ on R2R				SR ↑ on RxR-en	
			EnvDrop	FAST	Recur	PVLT	CLIPViL	HAMT
1	-	Vanilla	49.77	63.90	53.30	57.13	40.21	52.52
2	Object	MAV	-34%	-67%	-37%	-47%	-30%	-43%
3		MFG	-3%	-6%	-1%	-8%	-2%	-2%
4		CT	-5%	-10%	-6%	-9%	-3%	-5%
5	Direction	Flip	-41%	-30%	-48%	-59%	-36%	-47%

Table 10: Indoor navigation performance on various ablation settings on the visual environment. We compare three masking settings on the visual objects: mask all visible objects (MAV), mask foreground objects (MFG), and the controlled trial (CT). We horizontally flip the viewpoint to ablate direction-related visual information. Recur: VLN ⟳ BERT. PVLT: PREVALENT.

Dataset	Visual Feature	Vanilla	MAV	MFG	CT
R2R	ResNet-152	49.77	-34%	-3%	-5%
	CLIP-ViT	56.36	-34%	-2%	-5%
RxR-en	ResNet-152	35.27	-42%	-1%	-3%
	CLIP-ViT	40.21	-30%	-2%	-3%

Table 11: EnvDrop's navigation performance on R2R and RxR-en with different visual object ablation settings when using different visual features. MAV: mask all visible objects. MFG: mask foreground objects. CT: controlled trial.

out all the visible objects has a more salient impact on the success rate for all the listed indoor agents. We study the influence of masking visual objects at different depths by comparing the "mask foreground" setting with the controlled trial (#3 vs. #4). Noted here that the number of foreground objects is limited. Thus only a few objects are being masked out in both settings. Still, all listed indoor agents display worse performance on the controlled trial. Such results state that masking out further visual instances in the background, even only a tiny amount, will hurt navigation performance. This indicates that the tested agents consider all the objects in the visual environment during navigation, instead of merely staring at the closely surrounding objects.

Notice that the agent designs, the dataset domains, and the visual feature extractors are three coherent factors that may result in performance differences. We further justify this by adding another set of ablation studies, where we apply ImageNet ResNet-152 and CLIP-ViT to extract visual features for R2R and RxR-en, and evaluate with the same agent model EnvDrop. Results are shown in Table 11. The trend of different masking settings aligns with our previous findings in Table 10, and verifies that the background information is also crucial in the visual features.

5.2 Effect of Directions in the Environment

In this ablation setting, we randomly flip some of the viewpoints horizontally. The objects' relative positions at the flipped viewpoints will be reversed. Presumably, suppose the agent can follow the instruction and find the corresponding direction to approach. In that case, the flipped viewpoints will misguide the agent in the opposite direction and lower the navigation success rate. As shown in Table 10, flipping the viewpoints leads to drastic declines in the success rate for all listed indoor agents. This verifies our previous finding that indoor agents can understand directions in the instruction. We notice that FAST is the only listed model that is less affected by the direction flipping ablation than by the object masking ablation (#2 vs. #5). This suggests that FAST's asynchronous backtracking search is able to adjust and recover from errors that occur when choosing directions to some extent.

6 Analysis on Vision-Language Alignment

This section examines the agents' ability to learn vision-language alignment when executing the navigation. We focus on whether the agents can understand the objects mentioned in the instruction and align them to the correct visual instance in the environment, which is crucial to completing this multimodal task. To verify the existence of vision-language alignment, we add perturbations to the visual and textual input, and check how they affect agents' performance.

6.1 Instruction Side Perturbation

We add noise to the textual input by randomly replacing object tokens with random object tokens in the instruction. Table 5 shows an example (#3). This experiment aims to verify whether the agent can line the object tokens up to certain visual targets. The assumption is that if the agent can correctly align objects mentioned in the instruction to some targets in the visual environment, then replacing the object token will confuse and misguide the agent. Examining Figure 2, we notice that for all three datasets, the Transformer-based models have worse performance when replacing the object tokens, compared to simple masking. This indicates that Transformer-based models have a better cross-modal understanding of objects, and can align object tokens to the visual targets. Such superior performance may result from the fact that the Transformer-based models are often pre-trained on multimodal resources, thus displaying a slightly more vital ability to form alignment.

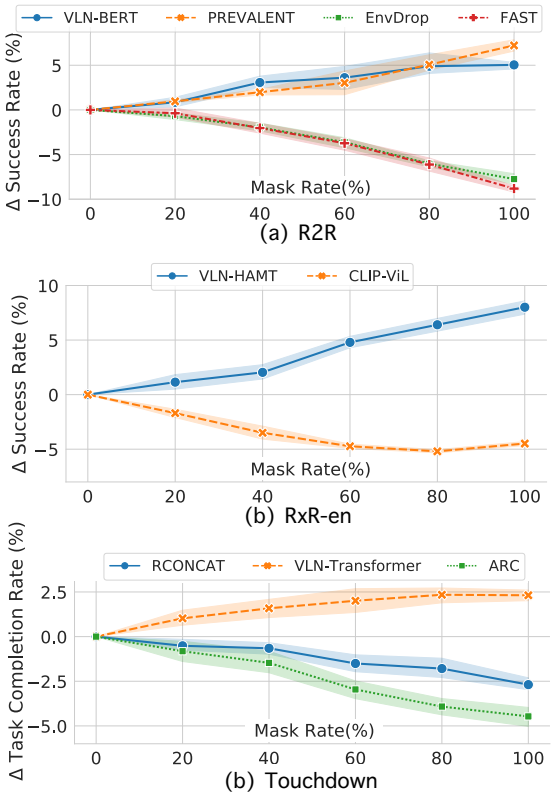

Figure 2: Performance gap between masking and replacing object tokens from instructions. If $\Delta > 0$, then replacing object tokens leads to worse navigation performance, which suggests that the agent has a better understanding of the object tokens.

Setting	SR ↑ on R2R				SR ↑ on RxR-en	
	EnvDrop	FAST	Recur	PVLT	CLIPViL	HAMT
Vanilla	49.77	63.90	53.30	57.13	40.21	52.52
Dynamic Mask	-10%	-24%	-5%	-13%	-19%	-23%
Controlled Trial	-8%	-19%	-3%	-11%	-16%	-20%
Mask Tokens	-36%	-38%	-21%	-20%	-48%	-32%

Table 12: Indoor navigation performance when dynamically masking out the visual objects mentioned in the instructions. Recur: VLN ⟳ BERT. PVLT: PREVALENT.

6.2 Environment Side Perturbation

We add noise to the visual input by conducting the following ablations. In the "dynamic mask" setting, we dynamically mask out the visual object regions mentioned in the instruction. We randomly mask out the same amount of visual objects at each viewpoint in its controlled trial. We also compare with the "mask tokens" setting, where we mask out all the object tokens in the instruction, while leaving the visual environment untouched. This

experiment aims to determine if the agent aligns the textual object tokens to the correct visual target. The assumption is that if the agent builds proper vision-language alignment and we mask out visual objects mentioned in the instruction, then the agent may get confused since it can not find the counterpart in the visual environment.

Results are shown in Table 12. The success rate witnesses a decline when dynamically masking out the visual objects. However, we notice in surprise that when all visual objects mentioned in the instruction are masked out, the agents can still reach a success rate higher than 44% on R2R and higher than 32% on RxR-en. This contradicts the previous assumption and casts doubt on the reliability of the navigation agents' vision-language alignment.

Comparing "dynamic mask" with the "mask tokens" setting, we notice that the visual object ablation has much smaller impact on navigation performance than the text object ablations, which suggests that current models have unbalanced attention on vision and text for the VLN task. Recent studies on pre-trained vision-and-language models (Frank et al., 2021) reveal that such asymmetry is also witnessed in other multimodal tasks. Future studies may follow the line of constructing a more balanced VLN agent.

7 Conclusion

In this paper, we inspect how the navigation agents understand the multimodal information by conducting ablation diagnostics input data. We find out that indoor navigation agents refer to both object tokens and direction tokens in the instruction when making decisions. In contrast, outdoor navigation agents heavily rely on direction tokens and poorly understand the object tokens. When it comes to vision-and-language alignments, we witness unbalanced attention on text and vision, and doubt the reliability of cross-modal alignments. We hope this work encourages more investigation and research into understanding neural VLN agents' black-box and improves the task setups and navigation agents' capacity for future studies.

Acknowledgement

We would like to show our gratitude to the anonymous reviewers for their thought-provoking comments. We would like to thank the Robert N. Noyce Trust for their generous gift to the University of California via the Noyce Initiative. The UCSB authors were also sponsored by the U.S. Army Research Office, and this work was accomplished under Contract Number W911NF19-D-0001 for the Institute for Collaborative Biotechnologies. The views and conclusions contained in this research are those of the authors and should not be interpreted as representing the sponsors or the U.S. government's official policy, expressed or inferred. Regardless of any copyright notation herein, the United States government is authorized to reproduce and distribute reprints for government purposes.

References

Anne H. Anderson, Miles Bader, Ellen Gurman Bard, Elizabeth Boyle, Gwyneth Doherty, Simon Garrod, Stephen D. Isard, Jacqueline C. Kowtko, Jan McAllister, Jim Miller, Catherine Sotillo, Henry S. Thompson, and Regina Weinert. 1991. The hcrc map task corpus. *Language and Speech*, 34:351 – 366.

Peter Anderson, Qi Wu, Damien Teney, Jake Bruce, Mark Johnson, Niko Sünderhauf, Ian D. Reid, Stephen Gould, and Anton van den Hengel. 2018. Vision-and-language navigation: Interpreting visually-grounded navigation instructions in real environments. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pages 3674–3683. IEEE Computer Society.

Jacob Andreas and Dan Klein. 2015. Alignment-based compositional semantics for instruction following. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1165–1174, Lisbon, Portugal. Association for Computational Linguistics.

Giuliano Antoniol, Roldano Cattoni, and Mauro Cettolo. 2011. Robust speech understanding for robot telecontrol.

Yoav Artzi and Luke Zettlemoyer. 2013. Weakly supervised learning of semantic parsers for mapping instructions to actions. *Transactions of the Association for Computational Linguistics*, 1:49–62.

Yonatan Bisk, Deniz Yuret, and Daniel Marcu. 2016a. Natural language communication with robots. In *NAACL*.

Yonatan Bisk, Deniz Yuret, and Daniel Marcu. 2016b. Natural language communication with robots. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 751–761, San Diego, California. Association for Computational Linguistics.

S.R.K. Branavan, Harr Chen, Luke Zettlemoyer, and Regina Barzilay. 2009. Reinforcement learning for mapping instructions to actions. In *Proceedings of*

the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP, pages 82–90, Suntec, Singapore. Association for Computational Linguistics.

S.R.K. Branavan, Luke Zettlemoyer, and Regina Barzilay. 2010. Reading between the lines: Learning to map high-level instructions to commands. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 1268–1277, Uppsala, Sweden. Association for Computational Linguistics.

Angel X. Chang, Angela Dai, Thomas A. Funkhouser, Maciej Halber, Matthias Nießner, Manolis Savva, Shuran Song, Andy Zeng, and Yinda Zhang. 2017. Matterport3d: Learning from RGB-D data in indoor environments. In *2017 International Conference on 3D Vision, 3DV 2017, Qingdao, China, October 10-12, 2017*, pages 667–676. IEEE Computer Society.

David L. Chen and Raymond J. Mooney. 2011. Learning to interpret natural language navigation instructions from observations. In *AAAI 2011*.

Howard Chen, Alane Suhr, Dipendra Misra, Noah Snavely, and Yoav Artzi. 2019. TOUCHDOWN: natural language navigation and spatial reasoning in visual street environments. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 12538–12547. Computer Vision Foundation / IEEE.

Shizhe Chen, Pierre-Louis Guhur, Cordelia Schmid, and Ivan Laptev. 2021. History aware multimodal transformer for vision-and-language navigation. In *NeurIPS*.

Soham Dan, Michael Zhou, and Dan Roth. 2021. Generalization in instruction following systems. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 976–981, Online. Association for Computational Linguistics.

Patrick Fernandes, Kayo Yin, Graham Neubig, and André F. T. Martins. 2021. Measuring and increasing context usage in context-aware machine translation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6467–6478, Online. Association for Computational Linguistics.

Stella Frank, Emanuele Bugliarello, and Desmond Elliott. 2021. Vision-and-language or vision-for-language? on cross-modal influence in multimodal transformers. In *EMNLP*.

Daniel Fried, Ronghang Hu, Volkan Cirik, Anna Rohrbach, Jacob Andreas, Louis-Philippe Morency, Taylor Berg-Kirkpatrick, Kate Saenko, Dan Klein, and Trevor Darrell. 2018. Speaker-follower models for vision-and-language navigation. In *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada*, pages 3318–3329.

Sergio Guadarrama, Lorenzo Riano, David Hamilton Golland, Daniel Goehring, Yangqing Jia, Dan Klein, P. Abbeel, and Trevor Darrell. 2013. Grounding spatial relations for human-robot interaction. *2013 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 1640–1647.

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 *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pages 13134–13143. IEEE.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 770–778. IEEE Computer Society.

Lisa Anne Hendricks, Zeynep Akata, Marcus Rohrbach, Jeff Donahue, Bernt Schiele, and Trevor Darrell. 2016. Generating visual explanations. In *Computer Vision - ECCV 2016 - 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part IV*, volume 9908 of *Lecture Notes in Computer Science*, pages 3–19. Springer.

Yicong Hong, Cristian Rodriguez Opazo, Yuankai Qi, Qi Wu, and Stephen Gould. 2020a. Language and visual entity relationship graph for agent navigation. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Yicong Hong, Qi Wu, Yuankai Qi, Cristian Rodriguez Opazo, and Stephen Gould. 2020b. A recurrent vision-and-language BERT for navigation. *CoRR*, abs/2011.13922.

Haoshuo Huang, Vihan Jain, Harsh Mehta, Alexander Ku, Gabriel Magalhães, Jason Baldridge, and Eugene Ie. 2019. Transferable representation learning in vision-and-language navigation. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*, pages 7403–7412. IEEE.

Vihan Jain, Gabriel Magalhães, Alexander Ku, Ashish Vaswani, Eugene Ie, and Jason Baldridge. 2019a. Stay on the path: Instruction fidelity in vision-and-language navigation. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 1862–1872. Association for Computational Linguistics.

Vihan Jain, Gabriel Magalhães, Alexander Ku, Ashish Vaswani, Eugene Ie, and Jason Baldridge. 2019b. Stay on the path: Instruction fidelity in vision-and-language navigation. *ArXiv*, abs/1905.12255.

Liyiming Ke, Xiujun Li, Yonatan Bisk, Ari Holtzman, Zhe Gan, Jingjing Liu, Jianfeng Gao, Yejin Choi, and Siddhartha S. Srinivasa. 2019. Tactical rewind: Self-correction via backtracking in vision-and-language navigation. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 6741–6749. Computer Vision Foundation / IEEE.

Volker Klingspor, John Demiris, and Michael Kaiser. 1997. Human-robot-communication and machine learning. *APPLIED ARTIFICIAL INTELLIGENCE JOURNAL*, 11(11):719–746.

Thomas Kollar, Jayant Krishnamurthy, and Grant P. Strimel. 2013. Toward interactive grounded language acqusition. In *Robotics: Science and Systems*.

Alexander Ku, Peter Anderson, Roma Patel, Eugene Ie, and Jason Baldridge. 2020a. Room-across-room: Multilingual vision-and-language navigation with dense spatiotemporal grounding. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 4392–4412. Association for Computational Linguistics.

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

Qing Li, Jianlong Fu, Dongfei Yu, Tao Mei, and Jiebo Luo. 2018. Tell-and-answer: Towards explainable visual question answering using attributes and captions. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*, pages 1338–1346. Association for Computational Linguistics.

Xiujun Li, Chunyuan Li, Qiaolin Xia, Yonatan Bisk, Asli Celikyilmaz, Jianfeng Gao, Noah A. Smith, and Yejin Choi. 2019. 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*, pages 1494–1499. Association for Computational Linguistics.

Jiasen Lu, Jianwei Yang, Dhruv Batra, and Devi Parikh. 2016. Hierarchical question-image co-attention for visual question answering. In *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain*, pages 289–297.

Chih-Yao Ma, Jiasen Lu, Zuxuan Wu, Ghassan AlRegib, Zsolt Kira, Richard Socher, and Caiming Xiong. 2019a. Self-monitoring navigation agent via auxiliary progress estimation. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.

Chih-Yao Ma, Zuxuan Wu, Ghassan AlRegib, Caiming Xiong, and Zsolt Kira. 2019b. The regretful agent: Heuristic-aided navigation through progress estimation. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 6732–6740. Computer Vision Foundation / IEEE.

Cynthia Matuszek, Liefeng Bo, Luke Zettlemoyer, and Dieter Fox. 2014. Learning from unscripted deictic gesture and language for human-robot interactions. In *AAAI*.

Harsh Mehta, Yoav Artzi, Jason Baldridge, Eugene Ie, and Piotr Mirowski. 2020. Retouchdown: Adding touchdown to streetlearn as a shareable resource for language grounding tasks in street view. *CoRR*, abs/2001.03671.

Hongyuan Mei, Mohit Bansal, and Matthew R. Walter. 2016. Listen, attend, and walk: Neural mapping of navigational instructions to action sequences. In *AAAI*.

Piotr Mirowski, Matthew Koichi Grimes, Mateusz Malinowski, Karl Moritz Hermann, Keith Anderson, Denis Teplyashin, Karen Simonyan, Koray Kavukcuoglu, Andrew Zisserman, and Raia Hadsell. 2018. Learning to navigate in cities without a map. In *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada*, pages 2424–2435.

Dipendra Misra, Andrew Bennett, Valts Blukis, Eyvind Niklasson, Max Shatkhin, and Yoav Artzi. 2018. Mapping instructions to actions in 3d environments with visual goal prediction. In *EMNLP*.

Dipendra Kumar Misra, John Langford, and Yoav Artzi. 2017. Mapping instructions and visual observations to actions with reinforcement learning. In *EMNLP*.

Khanh Nguyen, Debadeepta Dey, Chris Brockett, and Bill Dolan. 2019. Vision-based navigation with language-based assistance via imitation learning with indirect intervention. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 12527–12537. Computer Vision Foundation / IEEE.

Khanh Nguyen and Hal Daumé III. 2019. Help, anna! visual navigation with natural multimodal assistance via retrospective curiosity-encouraging imitation learning. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP*

2019, Hong Kong, China, November 3-7, 2019, pages 684–695. Association for Computational Linguistics.

Joe O'Connor and Jacob Andreas. 2021. What context features can transformer language models use? In *ACL/IJCNLP*.

Dong Huk Park, Lisa Anne Hendricks, Zeynep Akata, Anna Rohrbach, Bernt Schiele, Trevor Darrell, and Marcus Rohrbach. 2018. Multimodal explanations: Justifying decisions and pointing to the evidence. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pages 8779–8788. IEEE Computer Society.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. 2020a. Stanza: A python natural language processing toolkit for many human languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations, ACL 2020, Online, July 5-10, 2020*, pages 101–108. Association for Computational Linguistics.

Yuankai Qi, Zizheng Pan, Shengping Zhang, Anton van den Hengel, and Qi Wu. 2020b. Object-and-action aware model for visual language navigation. In *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part X*, volume 12355 of *Lecture Notes in Computer Science*, pages 303–317. Springer.

Yuankai Qi, Qi Wu, Peter Anderson, Xin Wang, William Yang Wang, Chunhua Shen, and Anton van den Hengel. 2020c. REVERIE: remote embodied visual referring expression in real indoor environments. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pages 9979–9988. IEEE.

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. 2021. Learning transferable visual models from natural language supervision. In *ICML*.

Deb K. Roy. 2002. Learning visually grounded words and syntax for a scene description task. *Comput. Speech Lang.*, 16:353–385.

Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. 2017. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*, pages 618–626. IEEE Computer Society.

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

Luc L. Steels and Paul Vogt. 1997. Grounding adaptive language games in robotic agents.

Christian Szegedy, Sergey Ioffe, Vincent Vanhoucke, and Alexander Amir Alemi. 2017. Inception-v4, inception-resnet and the impact of residual connections on learning. In *AAAI*.

Hao Tan and Mohit Bansal. 2018. Source-target inference models for spatial instruction understanding. In *AAAI*.

Hao Tan, Licheng Yu, and Mohit Bansal. 2019. Learning to navigate unseen environments: Back translation with environmental dropout. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 2610–2621. Association for Computational Linguistics.

Jesse Thomason, Daniel Gordon, and Yonatan Bisk. 2019. Shifting the baseline: Single modality performance on visual navigation & qa. In *NAACL*.

Adam Vogel and Daniel Jurafsky. 2010. Learning to follow navigational directions. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 806–814, Uppsala, Sweden. Association for Computational Linguistics.

Hu Wang, Qi Wu, and Chunhua Shen. 2020a. Soft expert reward learning for vision-and-language navigation. In *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part IX*, volume 12354 of *Lecture Notes in Computer Science*, pages 126–141. Springer.

Xin Wang, Qiuyuan Huang, Asli Celikyilmaz, Jianfeng Gao, Dinghan Shen, Yuan-Fang Wang, William Yang Wang, and Lei Zhang. 2019. Reinforced cross-modal matching and self-supervised imitation learning for vision-language navigation. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 6629–6638. Computer Vision Foundation / IEEE.

Xin Wang, Wenhan Xiong, Hongmin Wang, and William Yang Wang. 2018. Look before you leap: Bridging model-free and model-based reinforcement learning for planned-ahead vision-and-language navigation. In *Computer Vision - ECCV 2018 - 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part XVI*, volume 11220 of *Lecture Notes in Computer Science*, pages 38–55. Springer.

Xin Eric Wang, Vihan Jain, Eugene Ie, William Yang Wang, Zornitsa Kozareva, and Sujith Ravi. 2020b. Environment-agnostic multitask learning for natural language grounded navigation. In *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part XXIV*, volume 12369 of *Lecture Notes in Computer Science*, pages 413–430. Springer.

Terry Winograd. 1971. Procedures as a representation for data in a computer program for understanding natural language.

Jialin Wu and Raymond Mooney. 2019. Faithful multimodal explanation for visual question answering. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 103–112, Florence, Italy. Association for Computational Linguistics.

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

Jiannan Xiang, Xin Wang, and William Yang Wang. 2020. Learning to stop: A simple yet effective approach to urban vision-language navigation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings, EMNLP 2020, Online Event, 16-20 November 2020*, pages 699–707. Association for Computational Linguistics.

Fengda Zhu, Yi Zhu, Xiaojun Chang, and Xiaodan Liang. 2020a. Vision-language navigation with self-supervised auxiliary reasoning tasks. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pages 10009–10019. IEEE.

Wanrong Zhu, Xin Wang, Tsu-Jui Fu, An Yan, Pradyumna Narayana, Kazoo Sone, Sugato Basu, and William Yang Wang. 2020b. Multimodal text style transfer for outdoor vision-and-language navigation. *CoRR*, abs/2007.00229.

Yi Zhu, Fengda Zhu, Zhaohuan Zhan, Bingqian Lin, Jianbin Jiao, Xiaojun Chang, and Xiaodan Liang. 2020c. Vision-dialog navigation by exploring cross-modal memory. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pages 10727–10736. IEEE.