NDH-FULL: Learning and Evaluating Navigational Agents on Full-Length Dialogue

Hyounghun Kim **Jialu Li** **Mohit Bansal**
UNC Chapel Hill
{hyounghk, jialuli, mbansal}@cs.unc.edu

Abstract

Communication between human and mobile agents is getting increasingly important as such agents are widely deployed in our daily lives. Vision-and-Dialogue Navigation is one of the tasks that evaluate the agent's ability to interact with humans for assistance and navigate based on natural language responses. In this paper, we explore the Navigation from Dialogue History (NDH) task, which is based on the Cooperative Vision-and-Dialogue Navigation (CVDN) dataset, and present a state-of-the-art model which is built upon Vision-Language transformers. However, despite achieving competitive performance, we find that the agent in the NDH task is not evaluated appropriately by the primary metric – Goal Progress. By analyzing the performance mismatch between Goal Progress and other metrics (e.g., normalized Dynamic Time Warping) from our state-of-the-art model, we show that NDH's sub-path based task setup (i.e., navigating partial trajectory based on its correspondent subset of the full dialogue) does not provide the agent with enough supervision signal towards the goal region. Therefore, we propose a new task setup called NDH-FULL which takes the full dialogue and the whole navigation path as one instance. We present a strong baseline model and show initial results on this new task. We further describe several approaches that we try, in order to improve the model performance (based on curriculum learning, pre-training, and data-augmentation), suggesting potential useful training methods on this new NDH-FULL task.[1]

1 Introduction

With the increased number of intelligent agents being deployed in our daily lives, effective communication between humans and agents is becoming more important. Natural language is one of the most effective ways of communication due to its flexibility. Therefore, many efforts have been devoted to exploring the potential of its application in several tasks. Vision-and-Language Navigation (VLN) is one of the tasks in which agents have to navigate to a goal location in the indoor or outdoor environment by following natural language instructions (MacMahon et al., 2006; Tellex et al., 2011; Mei et al., 2016; Hermann et al., 2017; Brahmbhatt and Hays, 2017; Mirowski et al., 2018; Anderson et al., 2018; Misra et al., 2018; Blukis et al., 2019; Thomason et al., 2019; Nguyen and Daumé III, 2019; Chen et al., 2019; Shridhar et al., 2020; Qi et al., 2020; Hermann et al., 2020; Berg et al., 2020; Ku et al., 2020).

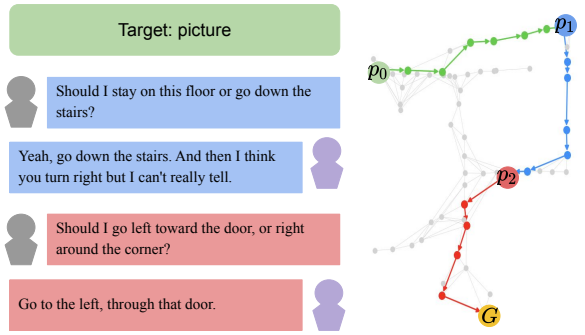


Figure 1: One example in the CVDN dataset. Given target information, dialogues in blue text and red text sequentially, the human navigates the green path, blue path, and red path accordingly.

While most VLN datasets only provide instructions from the oracle without considering the navigator's response, the useful Cooperative Vision-and-Dialogue Navigation (CVDN) (Thomason et al., 2019) dataset extends this one-way communication to two-way multi-turn dialogue (English) interaction between the oracle and the navigator. The dataset simulates a situation in which agents navigate through indoor environments towards a goal region by holding a conversation with humans for oracle guidance. Figure 1 shows an example in the CVDN dataset. Given the target in-

[1]Our code and dataset are publicly available at: https://github.com/hyounghk/NDH-FULL

Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pages 6432–6442
November 7–11, 2021.

formation *"picture"* only, the navigator is asked to explore the environment by intuition (green path). The navigator can ask the oracle for assistance during navigation and then make progress (blue and red path) based on the oracle's response. From this dataset, Thomason et al. (2019) proposed the Navigation from Dialogue History (NDH) task, in which the agents are asked to navigate toward the goal region G given dialogue history and the current round of the dialogue. However, we find that this sub-path-based task setup does not provide enough supervision for the agent to reach the goal region G, and its primary evaluation metric – Goal Progress (GP) does not appropriately measure the agent's performance on the sub-path based task. In the example shown in Figure 1, one CVDN example is split into three navigation instances starting from p_0, p_1, p_2 and ends at p_1, p_2, G, respectively. One NDH instance only contains dialogue before the current navigation path (e.g., for navigation from p_1 to p_2, the agent only knows the target *"picture"* and the first round of the dialogue, which is in the blue box), thus lacks supervision for how to navigate from p_2 to the goal region G. However, the agent is evaluated with GP – the distance made towards the goal region G from its starting point. This metric does not consider whether the agent follows the reference path. As a result, the agent could wander around to get a high GP score without following the path.

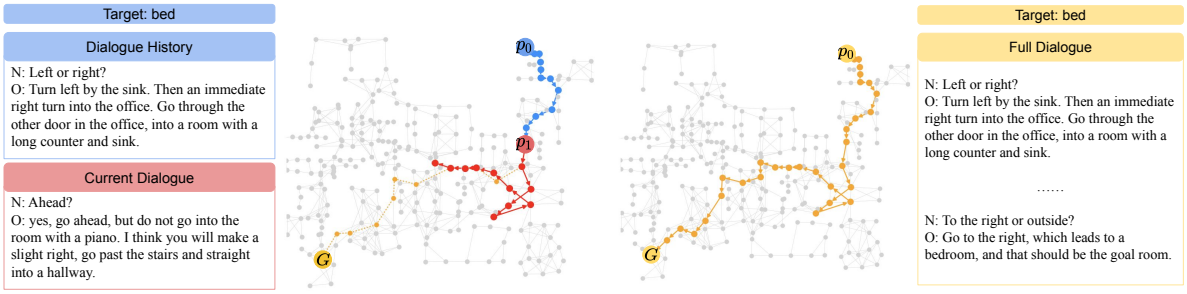
Hence, in this paper, we aim to redefine the NDH task via enhanced levels of supervision given to the agent, for better path fidelity while maintaining the advantage of learning from interactive dialogues. For this, we first build a strong state-of-the-art model based on Vision-Language transformers and pre-training, and illustrate that the current NDH task setup is not suitable for evaluating the agent's ability to follow natural language instructions. We show this by comparing the behaviors of the model on different evaluation metrics. Specifically, we find that a model with a higher GP score has a lower nDTW (normalized Dynamic Time Warping; Ilharco et al. (2019)) scores (see Table 3). Considering a high nDTW score reflects better path fidelity (and vice versa), pursuing high GP scores might not be suitable as an objective of an instruction-following navigation task. We attribute this mismatch to the aforementioned sub-path based task setup. Even though agents in the task could learn to navigate towards the target by commonsense and intuition, it might be hard to expect the agents to find the exact location of the target by using only their intuition (since this is hard even for human), especially in unseen environments since there is no specific regularity for target object placement (see Sec. 6.2 for analysis).

Therefore, we next propose a new task setup called NDH-FULL. We combine the sub-paths from the NDH task into the full path with the corresponding full dialogue, allowing the full supervision for agents on the instruction-following navigation task setup. As shown in the example of Figure 2, the NDH-FULL instance requires the agent to navigate from p_0 to G with full dialogue instruction (i.e., target and multiple rounds of dialogues). In this setting, the agent has explicit supervision towards the goal region and is further faced with the challenge of understanding and grounding longer dialogues to navigate longer paths compared with the NDH task. We present a strong baseline model and several enhancement suggestions (based on curriculum learning, pre-training, and data-augmentation) for this task, and still leaves a large room for useful future work by the community on this challenging and realistic NDH-FULL task setup.

Our contributions are three-fold: (1) We first present a state-of-the-art model for the NDH task. (2) We then demonstrate that the NDH task setup lacks supervision for reaching the goal region and its primary evaluation metric does not capture the agent's path fidelity (via both qualitative and quantitative analysis). (3) Thus, we propose a new challenging and realistic task setup called NDH-FULL (along with strong baseline models), which provides full paths with the corresponding full dialogue; and enhances supervision to encourage path fidelity.

2 Related Work

Vision-Language/Vision-Dialogue Navigation. In Vision-and-Language Navigation tasks, robots/agents are given natural language instructions and follow them in the outdoor or indoor environment to navigate and perform given tasks (MacMahon et al., 2006; Mooney, 2008; Chen and Mooney, 2011; Tellex et al., 2011; Mei et al., 2016; Hermann et al., 2017; Brahmbhatt and Hays, 2017; Mirowski et al., 2018; Anderson et al., 2018; Misra et al., 2018; Blukis et al., 2018; Das et al., 2018; Cirik et al., 2018; de Vries et al., 2018; Blukis



(a) NDH task setup

(b) NDH-FULL task setup

Figure 2: Comparison between the NDH task setup and the NDH-FULL task setup. Each sub-path corresponds to the sub-dialogue with same color. Dotted orange line in NDH task setup indicates shortest path between p_1 and goal region G. N indicates Navigator and O indicates Oracle in the dialogue.

et al., 2019; Thomason et al., 2019; Nguyen et al., 2019; Nguyen and Daumé III, 2019; Chen et al., 2019; Jain et al., 2019; Shridhar et al., 2020; Qi et al., 2020; Hermann et al., 2020; Berg et al., 2020; Zhu et al., 2020a; Ku et al., 2020; Anderson et al., 2020). Especially, Jain et al. (2019) introduces a new dataset, called Room-for-Room by combining short paths from Room-to-Room (Anderson et al., 2018) for evaluating instruction fidelity. Vision-and-Dialogue Navigation extends the one-way instruction-following navigation to the two-way multi-round dialogue setup in which agents could ask oracle guidance when they are lost. However, the current NDH task setup, which is built from the CVDN dataset (Thomason et al., 2019), does not provide enough supervision for agents' learning and does not evaluate agents' ability to navigate according to instructions. Thus, for better learning and evaluation, we introduce NDH-FULL which has the full path-dialogue pairs and leads to a more realistic, challenging setup.

Vision-Language Pre-Training. There have been significant improvements in natural language processing applications since large-scale pre-training language models were introduced (Radford et al., 2018; Devlin et al., 2019). The trend has spread to vision-language applications (Sun et al., 2019; Lu et al., 2019; Tan and Bansal, 2019; Chen et al., 2020; Li et al., 2020). Recently, the pre-training approach has shown promising results in vision-and-language navigation tasks as well (Majumdar et al., 2020; Hao et al., 2020; Hong et al., 2021). Following this trend, we also apply pre-training for our model. Compared with previous work, we take a more direct and effective approach by designing a pre-training model that is similar to the main navigation model and directly use VLN task as the pre-training objective.

3 Dataset Background and Task Setup

In this section, we discuss the vision-and-dialogue navigation task (NDH). We first introduce the CVDN dataset, and then show the two main issues of NDH and propose a new setup, NDH-FULL.

3.1 Cooperative Vision-Dialogue Navigation

The Cooperative Vision-and-Dialogue Navigation (CVDN) dataset contains dialogues between an oracle and a navigator. The navigator needs to find the target by asking questions during navigation. The oracle has access to the optimal navigation paths towards the target and responds to the navigator's questions. Specifically, each instance in the CVDN dataset contains a target object t_0, the start point for navigation p_0, the house scan S, the goal region G where the target object is located in, multiple turns of utterances between the oracle and navigator, and the navigator's corresponding navigation trajectories after interacting with the oracle.

3.2 Navigation from Dialogue History (NDH)

NDH Overview. Based on the CVDN dataset, Thomason et al. (2019) defines the task of Navigation from Dialogue History (NDH). In the NDH task, the navigation path is the sub-path of the full navigation path in the CVDN dataset. As shown in Figure 2, the start point for this NDH instance is p_1. The dialogue before this start point is recorded as the dialogue history. The red path is what a human navigator traverses based on target information, dialogue history, the current round of the dialogue, and navigation history from p_0 to p_1. In NDH, the agent is asked to find the target located in the goal region G based on this given information.

Issues with NDH Task Setup. Though many works (Hao et al., 2020; Zhu et al., 2020b; Wang

Task	Split	# of Inst.	Avg. PL	Avg. DL
NDH	Train	4742	7.68	3.82
	Val-Seen	382	7.61	4.31
	Val-Unseen	907	7.10	3.48
	Test-Unseen	1384	-	3.69
	Total	7415	7.59	3.78
NDH-FULL	Train	1145	25.82	5.79
	Val-Unseen	260	22.28	5.36
	Test-Unseen	248	24.42	5.56
	Total	1653	25.05	5.69

Table 1: Data statistics comparison between NDH and NDH-FULL. The average length of path and dialogue of NDH-FULL is longer than NDH's, implying NDH-FULL is a more challenging task. Since the path in Test-Unseen split of NDH task is not publicly released, the total Avg. PL is calculated except it (Inst.: instances, PL: path length, DL: dialouge length).

et al., 2020; Zhang et al., 2020) have made great progress in finding the target, the NDH task setup still has a couple of issues. First, the NDH task asks the agent to find the target without providing enough supervision, which makes this task hard even for human to finish. One instance in NDH does not contain further dialogue turns. Thus, based on the information which is only limited to the oracle's response and no further following dialogue rounds, the navigator cannot reach the target even with human intuition about where the target might be in an unseen room environment. As shown in Figure 2, given target information, dialogue history, the current round of the dialogue, and navigation history, a human navigator can only traverse the red path, which is still far away from the goal region where the target locates.

Second, the NDH task uses Goal Progress (GP) as the main metric to evaluate the navigation agent, which does not encourage instruction following and is not appropriate for measuring the performance on sub-path based task. As shown in Figure 2, the shortest path between p_1 and G does not align with the human's navigation according to dialogue information. The agent that navigates the shortest path or randomly explores the environment without following the instruction is not penalized by the GP metric. We show in Section 6.2 that the agent trained with the objective to have a higher GP will wander in the environment with long path length to get a GP without following the instruction, and thus deviates a lot from the reference path. This contradicts with the main goal of Vision-and-Language Navigation tasks which is to navigate environments by understanding instructions and grounding them with visual observations.

3.3 New Task Setup: NDH-FULL

In this section, we introduce the new task setup, NDH-FULL, to address the aforementioned issues in the NDH task. We create the NDH-FULL using the full dialogue-path pairs in CVDN. In other words, we combine multiple NDH instances that correspond to the same dialogue into one instance. As shown in Figure 2, given the target and full dialogue, the agent is asked to navigate from the start point t_0 to the goal region G. We also keep the sub-dialogue-path alignment information in the dataset, which brings the possibility for the agent to learn from sub-instructions. The NDH-FULL task setup provides full supervision for the agent to navigate towards the goal region and encourages the agent to understand long interactive dialogue and navigate with fidelity.

After combining all the sub-paths and dialogue turns into a full-length path-dialogue pair, the NDH-FULL has 1653 dialogue instances. We split them into training, validation-unseen, and test-unseen sets. We do not include validation seen set in NDH-FULL since we care more about agents' generalizability to unseen environments. The training, validation-unseen, and test-unseen sets contain 1145, 260, 248 instances respectively. Each of them is from 47, 10, and 10 non-overlapped scans, which preserves the important property that the environments of evaluation splits are unseen from the training set. We show detailed statistical comparison between NDH and NDH-FULL in Table 1. On average, the paths and dialogues of NDH-FULL are much longer than those of NDH (25.05 vs. 7.59 for path length, and 5.69 vs. 3.78 for dialogue length), which indicates that the NDH-FULL task setup is more challenging than NDH, allowing useful future work from the community. Furthermore, compared with NDH, the NDH-FULL gives the agent full supervision on how to reach the target and encourages the agent to understand long instructions and navigate based on the instructions.

4 NDH and NDH-FULL Models

We present the NDH task model and NDH-FULL task model in this section. To be specific, the NDH task model is built based on the vision-and-language transformer. Similar to the previous works (Hao et al., 2020; Hong et al., 2021), we employ LXMERT (Tan and Bansal, 2019) as the base architecture (Figure 3). The NDH-FULL task model takes the same architecture and is

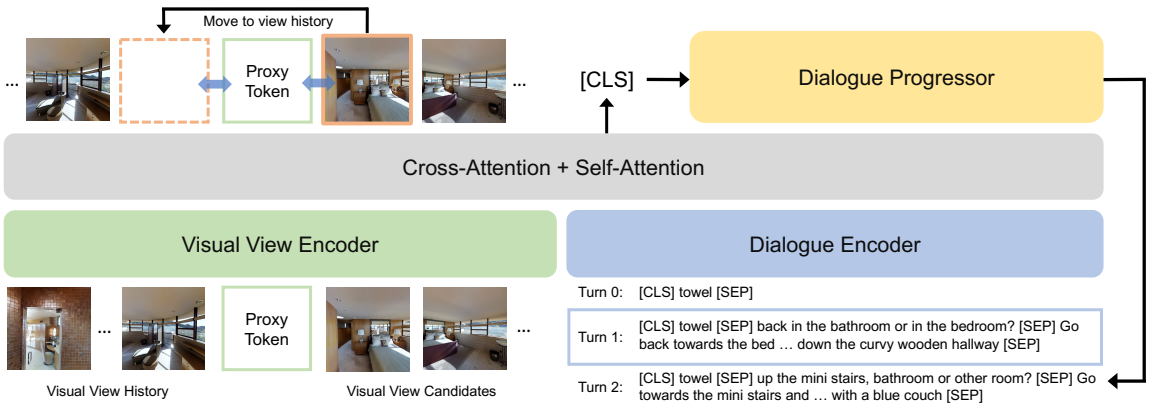


Figure 3: The dialogue navigation model on NDH-FULL task. The next view to proceed is selected based on the attention score between the visual proxy token and the candidate views. The dialogue progressor takes the current and next dialogue round features and decides whether to move to the next round or stay.

additionally equipped with the progressor module for moving through dialogue rounds. The NDH task model shows the state-of-the-art performance. However, by analyzing the behavior of the NDH task model on different metrics, we find the NDH task might not be suitable for evaluating the instructing-following navigation ability, thus, we propose the new NDH-FULL task and the baseline model (see Sec. 6.1 and 6.2).

Pre-Training Model. Pre-training is an effective approach to infuse prior knowledge in the vision-and-language navigation models (Majumdar et al., 2020; Hao et al., 2020; Hong et al., 2021). Compared with the previous works, our work proposes a new objective for pre-training. Instead of training the model with similarity score prediction (Majumdar et al., 2020) or discrete action label (Hao et al., 2020), we train the model with the objective that is nearly identical to the main navigation task for more effective transfer to the main task. Given a visual view sequence $V_t = \{v_1, v_2, ..., v_t\}$ and a corresponding navigation dialogue $D_i = \{d_{i0}, d_{i2}, ..., d_{i|D_i|}\}$, we train the model to select the next view to proceed among the candidates $C_t = \{c_{t1}, c_{t2}, ..., c_{t|C_t|}\}$. Additionally, we apply masked visual view prediction and masked language model loss as well. We employ ResNet (He et al., 2016) to get visual view features from panoramic images and use a multi-layer transformer to encode dialogue features like in LXMERT. The encoded features are fed to the LXMERT-based transformer module, TF_{LXT}.

$$L_n, L_v, L_l = \text{TF}_{LXT}([V_t^{\text{MASK}}; C_t], D_{1:i}^{\text{MASK}}) \quad (1)$$

where L_n, L_v, L_l are the losses for navigation task, masked visual view prediction, and masked language model, respectively. $[;]$ is the concatenation operation, $V_t^{\text{MASK}} = \{v_1, v_{[\text{MASK}]}, ..., v_{[\text{MASK}]}, ..., v_t\}$ and $D_i^{\text{MASK}} = \{d_{i0}, d_{i1}, ..., [\text{MASK}], d_{i|D_i|}\}$ are masked visual view and dialogue features, respectively. $D_{1:i}$ is concatenation of the dialogue features up to the ith round. To compute the navigation loss, we use multi-head attention score (of the last layer) between the current visual view v_t and the candidate visual views C_t as the action logit following Hong et al. (2021). TF_{LXT} consists of multiple layers of multi-head self-attention and cross-attention.

$$\hat{V}_t^j, \hat{D}_{1:i}^j = \text{MH-CrossATT}(V_t^j, D_{1:i}^j) \quad (2)$$

$$V_t^{j+1} = \text{MH-SelfATT}(\hat{V}_t^j) \quad (3)$$

$$D_{1:i}^{j+1} = \text{MH-SelfATT}(\hat{D}_{1:i}^j) \quad (4)$$

where MH-SelfATT is the multi-head self-attention and MH-CrossATT is the multi-head cross-attention. V_t^j and $D_{1:i}^j$ are the input of visual view and dialogue features to the jth layer, respectively. The lth self/cross attention head at jth layer is computed by (for the visual view feature case):

$$a_{j,l} = \text{Softmax}(\frac{QK^\top}{\sqrt{d_h}})V \quad (5)$$

$$Q = W_{j,l}^q C_t^{j-1},\ K = W_{j,l}^k V_t^{j-1},\ V = W_{j,l}^v V_t^{j-1} \quad (6)$$

$$V_t^j = [a_{j,1}; a_{j,2}; ...; a_{j,N_l}] \quad (7)$$

where $W_{j,l}^q$, $W_{j,l}^k$, and $W_{j,l}^v$ are trainable parameters, d_h is hidden dimension, and N_l is the number of attention heads. C_t^{j-1} can be V_t^{j-1} for self attention and $D_{1:i}^{j-1}$ for cross attention.

NDH Model. The dialogue navigation model for the NDH task shares the same base architecture as the pre-training model. On top of the pre-training model, we introduce the visual proxy token p_t which links the candidate views to the current and past view history (i.e., the candidate views and the current/past view history only communicate with the proxy token via attention, but they do not directly interact with each other). It also plays as the recurrent state feature which maintains context history information. By introducing the visual proxy token, the view candidates' logits are calculated from the multi-head attention scores between the visual proxy token and the view candidates. The visual proxy token allows the model to consider both explicit (past view history) and implicit (recurrent state) context.

$$\hat{c}_t, \hat{p}_t = \text{TF}_{LXT}([V_t; p_t; C_t], D_{1:i}) \quad (8)$$

$$p_{t+1} = \text{Linear}(\hat{p}_t) \quad (9)$$

where $\hat{c}_t$ is the predicted view to proceed. The visual proxy token of the last output layer from the TF_{LXT} model $\hat{p}_t$ is fed to a linear layer to become the visual proxy token at next time step.

NDH-FULL Model. For the NDH-FULL setup, we keep our strong NDH model as base architecture. In this model, we employ the CLIP visual feature (Radford et al., 2021) instead of the ResNet feature. To handle turns of the dialogue rounds, we introduce the dialogue progressor module which decides whether to move to the next round of the dialogue based on the current visual observation.

$$S_i = \text{Linear}([p_t; d_{i0}]) \quad (10)$$

$$S_{i+1} = \text{Linear}([p_t; d_{(i+1)0}]) \quad (11)$$

$$\text{NextTurn} = \begin{cases} i & \text{if } S_i \geq S_{i+1} \\ i+1 & \text{otherwise} \end{cases} \quad (12)$$

The dialogue progressor module simulates the situation in that the navigator is confused about which direction to go next and the oracle gives proper natural language guidance to the navigator. The progressor is trained from the alignment between sub-paths and corresponding dialogue rounds.

Mixture of Imitation and Reinforcement Learning. We use a mixture of imitation (IL) and reinforcement learning (RL) to train the model. For RL, we employ Actor-Critic (Mnih et al., 2016):

$$L_{IL} = -\sum_t \log p(a_t^*) \quad (13)$$

$$L_{RL} = -\sum_t (R_t - b_t) \log p(a_t^s) - \eta H(p(a_t)) \quad (14)$$

$$L_{MIX} = L_{IL} + \lambda L_{RL} \quad (15)$$

where R_t is the discounted cumulative reward, b_t is the baseline and $H(p(a_t))$ is the entropy term. a_t^* is the teacher action and a_t^s is the sampled action. We use distance-to-goal for the NHD task model and nDTW score for the NDH-FULL task model as the training rewards.

5 Experimental Setup

Metrics. We consider nDTW as the main metric of the new NDH-FULL task because nDTW reflects path fidelity better than other metrics (Ilharco et al., 2019). Other than nDTW, we also present evaluation results on success rate (SR), success weighted by path length (SPL), trajectory length (TL), and goal progress (GP) to allow evaluation from different perspectives.

Training Details. For the pre-training model, we use 9 language and 5 cross-modal LXMERT layers (but did not use their pre-trained weights), and use 768 as the hidden size. Following Tan and Bansal (2019), we use Adam (Kingma and Ba, 2015) as the optimizer with the learning rate 1×10^{-4} and linear decay as in Devlin et al. (2019). We use L2 loss for visual view prediction, and cross-entropy loss for masked language model and next view selection. We use CVDN (Thomason et al., 2019), R2R (Anderson et al., 2018), and a part of R2R's augmented data (Fried et al., 2018; Hao et al., 2020) as the training data. For the NDH task model, we use AdamW (Loshchilov and Hutter, 2018) as the optimizer with the learning rate 1×10^{-5}. Only CVDN data is used for fine-tuning the model. In the NDH-FULL task, we do not apply pre-training for the full-dialogue model. We use ResNet-152 feature and ResNet50-based CLIP feature. All the experiments are run using the NVIDIA TITAN Xp / GeForce GTX 1080 Ti / GeForce RTX 2080 Ti GPUs. We use PyTorch (Paszke et al., 2017) to build all models. We use manual tuning (e.g, learning rate=$\{1 \times 10^{-3}$, ..., $1 \times 10^{-6}\}$, and the layers of the transformer model=$\{$5(cross-modal)/3(language), 9/5$\}$) for selecting hyper-parameters. The number of trainable

Models	Val Unseen	Test Unseen
PREVALENT (Hao et al., 2020)	3.15	2.44
CMN (Zhu et al., 2020b)	2.97	3.14
EAML (Wang et al., 2020)	4.65	3.91
BabyWalk (Zhu et al., 2020a)	-	4.46
Ours	**5.51**	**5.27**

Table 2: Performance on NDH task measured with Goal Progress. Our model outperforms all the state-of-the-art models in the validation unseen environment and ranks 1st (at the time of EMNLP 2021 submission deadline) on the NDH task leaderboard ('s-agent' team). EAML: Environment-Agnostic Multitask Learning.

parameters of our NDH and NDH-FULL task models are 181M and 182M, respectively.

6 Results

6.1 State-of-the-Art Results on NDH Task

In this section, we present our model's performance on the NDH task. As shown in Table 2, our model outperforms all the state-of-the-art models on the primary evaluation metric – Goal Progress by a large margin and ranks 1st (at the time of EMNLP 2021 submission deadline) on the leaderboard ('s-agent' team).[2] This shows that our model performs strongly on the navigation task.

6.2 Analyzing the Issue in NDH Task Setup

However, we believe that the NDH task is not evaluated appropriately via the primary metric (i.e., GP) since GP could not reflect the instruction-following ability of the agents in the task. We conduct an experiment by running our model with two different rewards for reinforcement learning: global target reward and local target reward. In global target reward, the agent gets a positive reward if it moves closer to the final target region, and a negative reward otherwise. In local target reward, the agent receives the reward based on whether it moves closer to the final position of the sub-path. Since there is no explicit instruction for the path between the final position of the sub-path and the global target region (except when the sub-dialogue-path pair is the last pair in the full dialogue), the global target model stands for a model trained with implicit navigation supervision towards the global target region and the local target model stands for a model trained with no such implicit navigation supervision towards the global target region. We show the results in Table 3.

[2] https://eval.ai/web/challenges/challenge-page/463/leaderboard/1292

Reward Type	GP	SR	TL	nDTW	nDTW+
Global Target	5.51	19.8	24.582	0.253	0.243
Local Target	3.82	37.2	10.591	0.518	0.287

Table 3: Performance on NDH task (val-unseen split). Global Target is the model with a reward of distance to the final target and Local Target is the one with a reward of distance to the end point of the sup-path of each data instance in NDH task (nDTW+: nDTW based on extended reference path up to the target location).

Models	Val-Unseen				
	GP	SR	SPL	TL	nDTW
Random-Walk	5.755	3.1	2.8	10.056	0.141
No-Dialogue	10.972	6.5	5.6	29.556	0.267
Target-Only	10.005	6.2	4.9	29.828	0.267
Full-Dialogue	11.124	7.7	6.2	32.678	0.277
	Test-Unseen				
	14.045	10.5	7.6	28.539	0.301

Table 4: Performance on the new NDH-FULL task. The models are selected according to the best nDTW scores.

Goal Progress. The GP score of the global target model is much higher than the local target model (5.51 vs. 3.82), indicating that the global target model reaches closer to the global target location with implicit supervision.

Instruction Following. However, when we compare the success rate scores (19.8 vs. 37.2) and nDTW scores (0.253 vs. 0.518), the local target model outperforms the global target model, indicating that the local target model follows the reference path better. This mismatch in metrics implies that GP cannot measure the agent's ability to follow the path well.

Intuition to Reach Target. A higher GP score of the global target model can be considered as the result of learning intuition to navigate towards the target region without explicit supervision. However, we show in Table 3 that the global target model has a much higher trajectory length (TL) compared with the local target model (24.582 vs. 10.591), indicating that the agent learns to get a higher GP by wandering in the environment rather than proceeding towards a specific direction with intuition. We also show that the global target model has a lower nDTW+ score (which is a nDTW score against the extended reference path to the target location measuring the agent's ability to follow the path from the current starting point to the target) compared with the local target model (0.243 vs. 0.287), which also supports the observation that the global target

Models	Val-Unseen				
	GP	SR	SPL	TL	nDTW
Curriculum Learning	11.241	7.3	6.3	32.169	0.273
Pre-Training	12.268	7.3	6.2	34.166	0.278
Data Augmentation	11.058	6.9	5.9	35.306	0.263

Table 5: Performance from different approaches on the new NDH-FULL task.

model does not follow the extended path towards the global target region with intuition to get a high GP score.

Therefore, pursuing higher GP scores might not reflect agents' ability to interpret and follow given dialogues. For this reason, we introduce a new task setup, NDH-FULL, which encourages instruction following by giving full supervision towards the global target to the agent.

6.3 NDH-FULL Task Results & Suggestions

We show the performance of our model and its ablations on the new NDH-FULL task. We experiment with the "Random-Walk" baseline which chooses a random heading and walks up to 5 steps forward as in Thomason et al. (2019), "No-Dialogue" baseline which only considers visual input, and "Target-Only" baseline which considers visual input and the target information. As shown in Table 4, with full supervision towards the target goal region (Full-Dialogue), the agent outperforms the other baselines in all metrics, which indicates that full-dialogue provides useful supervision for the agent. However, performance gap between models is not large. Considering the full-dialogue model shows the best performance in the NDH task, the new NDH-FULL task is quite challenging with longer paths and dialogues. Moreover, requirement of aligning each sub-path and the corresponding dialogue round in the NDH-FULL task introduces additional dimension of difficulty to handle for better performance in instruction-following navigation.

Therefore, we believe there is still a large room for potential improvement by applying more advanced approaches. Thus, we experiment with some of the advanced approaches here as an initial step to tackle this challenge.

Curriculum Learning. We divide one data instance into multiple instances so that each resulting data point has a different number of dialogue rounds and a corresponding sub-path (i.e., 2, 3, and 4 or more than 4 dialogue rounds) and train the model on the subset of the data and move on to the longer dialogue/path ones (starting from the 2 dialogue rounds to the original full dialogue rounds). But, as shown in Table 5, this curriculum learning approach only does not show an improvement. With a more finely designed learning procedure, we believe curriculum learning would help improve the performance on the challenging new task.

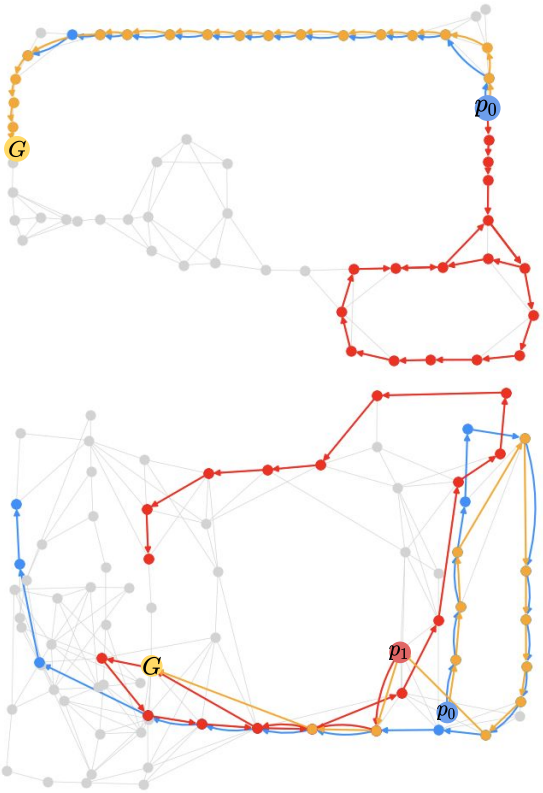


Figure 4: Trajectory comparison between the NDH task model (red line) and the NDH-FULL task model (blue line). Yellow line is the reference path (p_0: starting point of the whole path, p_1: starting point of 2nd sub-path, G: goal point).

Pre-Training. We also apply the pre-trained weights which are used for the NDH model. However, this also does not give any distinct performance boost. This might be because the pre-training model for the NDH task is passive in that the model is given visual and textual features at once. On the other hand, in the NDH-FULL task, agents should actively ask for guidance when they are confused. Therefore, aligning dialogue rounds with the visual observation from the environment is one challenging factor in the new task.

Data Augmentation. The data size of NDH-FULL shrinks after combining all sub-paths and dialogue rounds (7415 vs. 1653, see Table 1). To compensate for the loss, we try data augmentation by generating the oracle's instruction with the speaker model (Fried et al., 2018; Tan et al., 2019). We

modify their speaker model to take the context (i.e., dialogue history) as well as view trajectory to fit to the CVDN dataset. We replace the oracle's instruction in a round of dialogue with the newly generated ones to give the model more diverse forms of instructions. But, we do not see an improvement from training the model on this augmented data possibly because NDH-FULL requires accurate instructions to navigate quite long paths and the quality of the current speaker model could not meet the criteria. This allows future work on more effective generation methods.

6.4 Trajectory Comparison

As shown from the top figure in Figure 4, the NDH task agent (red line) fails to follow the correct reference trajectory (yellow line) by misunderstanding the oracle's instruction (*"turn around and follow the red carpet path. Once you pass a vase on your left stop"*) while still getting a positive GP score (8.820). On the other hand, the NDH-FULL task agent (blue line) can manage to follow the instructions showing a high path fidelity (nDTW score: 0.735). This example implies that GP is not a good metric for measuring instruction-following. In the bottom example, the NDH task agent starts from p_1 (in the sub-path task setup) and move towards the goal location, but it directly passes the target object and wanders in the room. This trajectory deviates much from the reference sub-path, but the agent still gets a high GP (8.226) since it finally stops near the goal region. Though the NDH-FULL task agent doesn't stop at the goal region either, it follows the reference path well during most of the navigation process (nDTW score: 0.549).

7 Conclusion

We explored the NDH task, which is built on the useful Cooperative Vision-and-Dialogue Navigation (CVDN) dataset, and found the mismatch between the task setup and evaluation by analyzing the scoring behaviors of our state-of-the-art model. Therefore, we proposed a new task called NDH-FULL. We combined all split paths and dialogue rounds of NDH to create the full path and dialogue, resulting NDH-FULL has longer paths and dialogues than NDH and it makes NDH-FULL more challenging. We also presented a baseline model, resulting scores, and suggestions for promising research directions on the NDH-FULL task.

Acknowledgments

We thank the reviewers for their helpful comments. This work was supported by NSF Award 1840131, ARO-YIP Award W911NF-18-1-0336, DARPA KAIROS Grant FA8750-19-2-1004, and a Google Focused Award. The views contained in this article are those of the authors and not of the funding agency.

References

Peter Anderson, Ayush Shrivastava, Joanne Truong, Arjun Majumdar, Devi Parikh, Dhruv Batra, and Stefan Lee. 2020. Sim-to-real transfer for vision-and-language navigation. *CoRL*.

Peter Anderson, Qi Wu, Damien Teney, Jake Bruce, Mark Johnson, Niko Sünderhauf, Ian Reid, Stephen Gould, and Anton van den Hengel. 2018. Vision-and-language navigation: Interpreting visually-grounded navigation instructions in real environments. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3674–3683.

Matthew Berg, Deniz Bayazit, Rebecca Mathew, Ariel Rotter-Aboyoun, Ellie Pavlick, and Stefanie Tellex. 2020. Grounding language to landmarks in arbitrary outdoor environments. In *2020 IEEE International Conference on Robotics and Automation (ICRA)*, pages 208–215. IEEE.

Valts Blukis, Dipendra Misra, Ross A Knepper, and Yoav Artzi. 2018. Mapping navigation instructions to continuous control actions with position-visitation prediction. In *Conference on Robot Learning*, pages 505–518.

Valts Blukis, Yannick Terme, Eyvind Niklasson, Ross A Knepper, and Yoav Artzi. 2019. Learning to map natural language instructions to physical quadcopter control using simulated flight. In *Conference on Robot Learning*, pages 1415–1438.

Samarth Brahmbhatt and James Hays. 2017. Deepnav: Learning to navigate large cities. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5193–5202.

David L Chen and Raymond J Mooney. 2011. Learning to interpret natural language navigation instructions from observations. In *Twenty-Fifth AAAI Conference on Artificial Intelligence*.

Howard Chen, Alane Suhr, Dipendra Misra, Noah Snavely, and Yoav Artzi. 2019. Touchdown: Natural language navigation and spatial reasoning in visual street environments. In *Conference on Computer Vision and Pattern Recognition*.

Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and

Jingjing Liu. 2020. Uniter: Universal image-text representation learning. In *European Conference on Computer Vision*, pages 104–120. Springer.

Volkan Cirik, Yuan Zhang, and Jason Baldridge. 2018. Following formulaic map instructions in a street simulation environment. In *2018 NeurIPS Workshop on Visually Grounded Interaction and Language*, volume 1.

Abhishek Das, Samyak Datta, Georgia Gkioxari, Stefan Lee, Devi Parikh, and Dhruv Batra. 2018. Embodied Question Answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

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

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *NAACL*.

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 *NeurIPS*.

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.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *CVPR*, pages 770–778.

Karl Moritz Hermann, Felix Hill, Simon Green, Fumin Wang, Ryan Faulkner, Hubert Soyer, David Szepesvari, Wojciech Marian Czarnecki, Max Jaderberg, Denis Teplyashin, et al. 2017. Grounded language learning in a simulated 3d world. *arXiv preprint arXiv:1706.06551*.

Karl Moritz Hermann, Mateusz Malinowski, Piotr Mirowski, Andras Banki-Horvath, Keith Anderson, and Raia Hadsell. 2020. Learning to follow directions in street view. *Thirty-Fourth AAAI Conference on Artificial Intelligence*.

Yicong Hong, Qi Wu, Yuankai Qi, Cristian Rodriguez-Opazo, and Stephen Gould. 2021. A recurrent vision-and-language bert for navigation. *CVPR*.

Gabriel Ilharco, Vihan Jain, Alexander Ku, Eugene Ie, and Jason Baldridge. 2019. Effective and general evaluation for instruction conditioned navigation using dynamic time warping. *NeurIPS Visually Grounded Interaction and Language Workshop*.

Vihan Jain, Gabriel Magalhaes, Alexander Ku, Ashish Vaswani, Eugene Ie, and Jason Baldridge. 2019. Stay on the Path: Instruction Fidelity in Vision-and-Language Navigation. In *Proc. of ACL*.

Diederik P Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *ICLR*.

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

Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, et al. 2020. Oscar: Object-semantics aligned pre-training for vision-language tasks. In *European Conference on Computer Vision*, pages 121–137. Springer.

Ilya Loshchilov and Frank Hutter. 2018. Decoupled weight decay regularization. In *International Conference on Learning Representations*.

Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *NeurIPS*.

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

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*, pages 259–274. Springer.

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

Piotr Mirowski, Matt Grimes, Mateusz Malinowski, Karl Moritz Hermann, Keith Anderson, Denis Teplyashin, Karen Simonyan, Andrew Zisserman, Raia Hadsell, et al. 2018. Learning to navigate in cities without a map. In *Advances in Neural Information Processing Systems*, pages 2419–2430.

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

Volodymyr Mnih, Adria Puigdomenech Badia, Mehdi Mirza, Alex Graves, Timothy Lillicrap, Tim Harley, David Silver, and Koray Kavukcuoglu. 2016. Asynchronous methods for deep reinforcement learning.

In *International conference on machine learning*, pages 1928–1937. PMLR.

Raymond J Mooney. 2008. Learning to connect language and perception. In *AAAI*, pages 1598–1601.

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 Conference on Empirical Methods in Natural Language Processing (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 *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. 2017. Automatic differentiation in PyTorch. In *NIPS Autodiff Workshop*.

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 Conference on Computer Vision and Pattern Recognition (CVPR)*.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. *arXiv preprint arXiv:2103.00020*.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving language understanding by generative pre-training.

Mohit Shridhar, Jesse Thomason, Daniel Gordon, Yonatan Bisk, Winson Han, Roozbeh Mottaghi, Luke Zettlemoyer, and Dieter Fox. 2020. ALFRED: A Benchmark for Interpreting Grounded Instructions for Everyday Tasks. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Chen Sun, Austin Myers, Carl Vondrick, Kevin Murphy, and Cordelia Schmid. 2019. Videobert: A joint model for video and language representation learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7464–7473.

Hao Tan and Mohit Bansal. 2019. Lxmert: Learning cross-modality encoder representations from transformers. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*.

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

Stefanie A Tellex, Thomas Fleming Kollar, Steven R Dickerson, Matthew R Walter, Ashis Banerjee, Seth Teller, and Nicholas Roy. 2011. Understanding natural language commands for robotic navigation and mobile manipulation. In *AAAI*.

Jesse Thomason, Michael Murray, Maya Cakmak, and Luke Zettlemoyer. 2019. Vision-and-dialog navigation. In *Conference on Robot Learning (CoRL)*.

Xin Wang, Vihan Jain, Eugene Ie, William Yang Wang, Zornitsa Kozareva, and Sujith Ravi. 2020. Environment-agnostic multitask learning for natural language grounded navigation. *ECCV*.

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

Wang Zhu, Hexiang Hu, Jiacheng Chen, Zhiwei Deng, Vihan Jain, E. Ie, and Fei Sha. 2020a. Babywalk: Going farther in vision-and-language navigation by taking baby steps. In *ACL*.

Yi Zhu, Fengda Zhu, Zhaohuan Zhan, Bingqian Lin, Jianbin Jiao, Xiaojun Chang, and Xiaodan Liang. 2020b. Vision-dialog navigation by exploring cross-modal memory. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10730–10739.