

# Towards Learning a Generic Agent for Vision-and-Language Navigation via Pre-training

Weituo Hao<sup>1†‡</sup>, Chunyuan Li<sup>2†\*</sup>, Xiujun Li<sup>2</sup>, Lawrence Carin<sup>1</sup>, Jianfeng Gao<sup>2</sup>  
<sup>1</sup>Duke University    <sup>2</sup>Microsoft Research, Redmond

{weituo.hao, lcarin}@duke.edu    {chunyl, xiul, jfgao}@microsoft.com

## Abstract

*Learning to navigate in a visual environment following natural-language instructions is a challenging task, because the multimodal inputs to the agent are highly variable, and the training data on a new task is often limited. We present the first pre-training and fine-tuning paradigm for vision-and-language navigation (VLN) tasks. By training on a large amount of image-text-action triplets in a self-supervised learning manner, the pre-trained model provides generic representations of visual environments and language instructions. It can be easily used as a drop-in for existing VLN frameworks, leading to the proposed agent PREVALENT<sup>1</sup>. It learns more effectively in new tasks and generalizes better in a previously unseen environment. The performance is validated on three VLN tasks. On the Room-to-Room [3] benchmark, our model improves the state-of-the-art from 47% to 51% on success rate weighted by path length. Further, the learned representation is transferable to other VLN tasks. On two recent tasks, vision-and-dialog navigation [30] and “Help, Anna!” [22], the proposed PREVALENT leads to significant improvement over existing methods, achieving a new state of the art.*

## 1. Introduction

Learning to navigate in a photorealistic home environment based on natural language instructions has attracted increasing research interest [23, 14, 7, 3, 6], as it provides insight into core scientific questions about multimodal representations. It also takes a step toward real-world applications, such as personal assistants and in-home robots. Vision-and-language navigation (VLN) presents a challenging reasoning problem for agents, as the multimodal inputs are highly variable, inherently ambiguous, and often under-specified.

\*Corresponding author †Equal Contribution ‡Work performed during an internship at MSR

<sup>1</sup>PRE-TRAINED VISION-AND-LANGUAGE BASED NAVIGATOR

Most previous methods build on the sequence-to-sequence architecture [26], where the instruction is encoded as a sequence of words, and the navigation trajectory is decoded as a sequence of actions, enhanced with attention mechanisms [3, 32, 18] and beam search [9]. While a number of methods [20, 21, 33] have been proposed to improve language understanding, common to all existing work is that the agent learns to understand each instruction from scratch or in isolation, without collectively leveraging prior vision-grounded domain knowledge.

However, each instruction in practice only loosely aligns with the desired navigation path, making it imperfect for the existing paradigm of learning to understand the instruction from scratch. This is because (i) every instruction only partially characterizes the trajectory. It can be ambiguous to interpret the instructions, without grounding on the visual states. (ii) The objects in visual states and language instructions may share various common forms/relationships, and therefore it is natural to build an informative joint representation beforehand, and use this “common knowledge” for transfer learning in downstream tasks.

To address this natural ambiguity of instructions more effectively, we propose to pre-train an encoder to align language instructions and visual states for joint representations. The image-text-action triplets at each time step are independently fed into the model, which is trained to predict the masked word tokens and next actions, thus formulating the VLN pre-training in the self-learning paradigm. The complexity of VLN learning can then be reduced by eliminating language understandings which lack consensus from visual states. The pre-trained model plays the role of providing generic image-text representations, and is applicable to most existing approaches to VLN, leading to our agent PREVALENT. We consider three VLN scenarios as downstream tasks: Room-to-room (R2R) [3], cooperative vision-and-dialog navigation (CVDN) [30], and “Help, Anna!” (HANNA) [22]. The overall pre-training and fine-tuning pipeline is shown in Figure 1.

Comprehensive experiments demonstrate strong empirical performance of PREVALENT. The proposed PREVA-

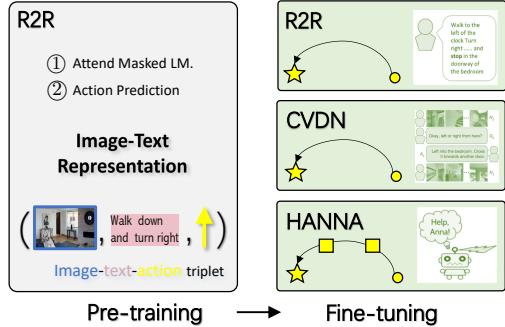


Figure 1: Illustration of the proposed pre-training and fine-tuning paradigm for VLN. The image-text-action triplets are collected from the R2R dataset. The model is pre-trained with two self-supervised learning objectives, and fine-tuned for three tasks: R2R, CVND and HANNA. R2R is an in-domain task, where the language instruction is given at the beginning, describing the full navigation path. CVND and HANNA are out-of-domain tasks; the former is to navigate based on dialog history, while the latter is an interactive environment, where intermediate instructions are given in the middle of navigation.

LENT achieves a new state of the art on all three tasks <sup>2</sup>. Comparing with existing methods, it adapts faster, and generalizes better to unseen environments and new tasks. Our code and pre-trained model is released on GitHub <sup>3</sup>.

## 2. Related Work

**Vision-language pre-training** Vision-Language Pre-trainig (VLP) is a rapidly growing research area. The existing approaches employ BERT-like objectives [8] to learn cross-modal representation for various vision-language problems, such as visual question-answering, image-text retrieval and image captioning *etc.* [25, 27, 17, 34, 24, 15]. However, these VLP works focus on learning representations only for vision-language domains. This paper presents the first pre-trained models, grounding vision-language understanding with actions in a reinforcement learning setting. Further, existing VLP methods require faster R-CNN features as visual inputs [10, 2], which are not readily applicable to VLN. State-of-the-art VLN systems are based on panoramic views (*e.g.*, 36 images per view for R2R), and therefore it is computationally infeasible to extract region features for all views and feed them into the agent.

**Vision-and-language navigation** Various methods have been proposed for learning to navigate based on vision-language cues. In [9] a panoramic action space and a

“speaker” model were introduced for data augmentation. A novel neural decoding scheme was proposed in [12] with search, to balance global and local information. To improve the alignment of the instruction and visual scenes, a visual-textual co-grounding attention mechanism was proposed in [18], which is further improved with a progress monitor [19]. To improve the generalization of the learned policy to unseen environments, reinforcement learning has been considered, including planning [33], and exploration of unseen environments using a off-policy method [32]. An environment dropout was proposed [28] to generate more environments based on the limited data, so that it can generalize well to unseen environments. These methods are specifically designed for particular tasks, and hard to generalize for new tasks. In this paper, we propose the first generic agent that is pre-trained to effectively understand vision-language inputs for a broad range of navigation tasks, and can quickly adapt to new tasks. The most related agent to ours is PRESS [16]. However, our work is different from [16] from two perspectives: (*i*) PRESS employs an off-the-shelf BERT [8] model for language instruction understanding, while we pre-train a vision-language encoder from scratch, specifically for the navigation tasks. (*ii*) PRESS only focuses on the R2R task, while we verify the effectiveness of our pre-trained model on three tasks, including two out-of-domain navigation tasks.

## 3. Background

The VLN task can be formulated as a Partially Observable Markov Decision Process (POMDP)  $\mathcal{M} = \langle \mathcal{S}, \mathcal{A}, P_s, r \rangle$ , where  $\mathcal{S}$  is the visual state space,  $\mathcal{A}$  is a discrete action space,  $P_s$  is the unknown environment distribution from which we draw the next state, and  $r \in \mathbb{R}$  is the reward function. At each time step  $t$ , the agent first observes an RGB image  $s_t \in \mathcal{S}$ , and then takes an action  $a_t \in \mathcal{A}$ . This leads the simulator to generate a new image observation  $s_{t+1} \sim P_s(\cdot | s_t, a_t)$  as the next state. The agent interacts with the environment sequentially, and generates a trajectory of length  $T$ . The episode ends when the agent selects the special STOP action, or when a pre-defined maximum trajectory length is reached. The navigation is successfully completed if the trajectory  $\tau$  terminates at the intended target location.

In a typical VLN setting, the instructions are represented as a set  $\mathcal{X} = \{\mathbf{x}_i\}_{i=1}^M$ , where  $M$  is the number of alternative instructions, and each instruction  $\mathbf{x}_i$  consists of a sequence of  $L_i$  word tokens,  $\mathbf{x}_i = [x_{i,1}, x_{i,2}, \dots, x_{i,L_i}]$ . The training dataset  $\mathcal{D}_E = \{\tau, \mathbf{x}\}$  consists of pairs of the instruction  $\mathbf{x}$  together with its corresponding expert trajectory  $\tau$ . The agent then learns to navigate via performing maximum likelihood estimation (MLE) of the policy  $\pi$ , based on the

<sup>2</sup>Among *all* public results at the time of this submission.

<sup>3</sup><https://github.com/weituo12321/PREVALENT>

individual sequences:

$$\max_{\theta} \mathcal{L}_{\theta}(\tau, \mathbf{x}) = \log \pi_{\theta}(\tau | \mathbf{x}) = \sum_{t=1}^T \log \pi_{\theta}(\mathbf{a}_t | \mathbf{s}_t, \mathbf{x}), \quad (1)$$

where  $\theta$  are the policy parameters. The policy is usually parameterized as an attention-based Seq2Seq model [3, 9], trained in the teacher-forcing fashion, *i.e.*, the ground-truth states  $\mathbf{s}_t$  are provided at every step in training. This allows reparameterization of the policy as an encoder-decoder architecture, by considering a function decomposition  $\pi_{\theta} = f_{\theta_E} \circ f_{\theta_D}$ :

- A *vision-language encoder*  $f_{\theta_E} : \{\mathbf{s}_t, \mathbf{x}\} \rightarrow \mathbf{z}_t$ , where a joint representation  $\mathbf{z}_t$  at time step  $t$  is learned over the visual state  $\mathbf{s}_t$  and the language instruction  $\mathbf{x}$ .
- An *action decoder*  $f_{\theta_D} : \{\mathbf{s}_t, \mathbf{z}_t\} \rightarrow \mathbf{a}_t$ . For each joint representation  $\mathbf{s}_t$ , we ground it with  $\mathbf{s}_t$  via neural attention, and decode into actions  $\mathbf{a}_t$ .

Successful navigation largely depends on precise joint understanding of natural language instructions and the visual states [29]. We isolate the encoder stage, and focus on pre-training a generic vision-language encoder for various navigation tasks.

## 4. Pre-training Models

Our pre-training model aims to provide joint representations for image-text inputs in VLN.

### 4.1. Input Embeddings

The input embedding layers convert the inputs (*i.e.*, panoramic views and language instruction) into two sequences of features: image-level visual embeddings and word-level sentence embeddings.

**Visual Embedding** Following [9], we employ panoramic views as visual inputs to the agent. Each panoramic view consists of 36 images in total (12 angles, and 3 camera poses per angle):  $\mathbf{s} = [s_1, \dots, s_{36}]$ . Each image is represented as a 2176-dimensional feature vector  $s = [s_v, s_p]$ , as a result of the concatenation of two vectors: (*i*) The 2048-dimensional visual feature  $s_v$  output by a Residual Network (ResNet) of the image [11]; (*ii*) the 128-dimensional orientation feature vector  $s_p$  that repeats  $[\sin \psi; \cos \psi; \sin \omega; \cos \omega]$  32 times, where  $\psi$  and  $\omega$  are the heading and elevation poses, respectively [9]. The embedding for each image is:

$$\mathbf{h} = \text{Layer-Norm}(\mathbf{W}_e s + \mathbf{b}_e) \quad (2)$$

where  $\mathbf{W}_e \in \mathbb{R}^{d_h \times 2176}$  is a weight matrix, and  $\mathbf{b}_e \in \mathbb{R}^{d_h}$  is the bias term;  $d_h = 768$  in our experiments. Layer normalization (LN) [4] is used on the output of this fully connected (FC) layer. An illustration of the visual embedding is shown in Figure 2(a).

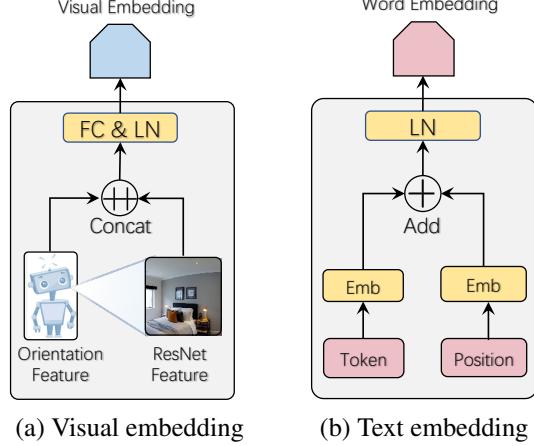


Figure 2: Illustration for the representation procedure of (a) visual embedding and (b) text embedding. FC is the fully-connected layer, and LN is the layer-normalization layer.

**Text Embedding** The embedding layer for the language instruction follows the standard Transformer, where LN is applied to the summation of the token embedding and position embedding. An illustration of the text embedding is shown in Figure 2(b).

### 4.2. Encoder Architecture

Our backbone network has three principal modules: two single-modal encoders (one for each modality), followed by a cross-modal encoder. All modules are based on a multi-layer Transformer [31]. For the  $\ell$ -th Transformer layer, its output is

$$\mathbf{H}_{\ell} = \mathcal{T}(\mathbf{H}_{\ell-1}, \mathbf{H}', \mathbf{M}) \quad (3)$$

where  $\mathbf{H}_{\ell-1} \in \mathbb{R}^{L \times d_h}$  is the previous layer's features ( $L$  is the sequence length),  $\mathbf{H}' \in \mathbb{R}^{L' \times d_h}$  is the feature matrix to attend, and  $\mathbf{M} \in \mathbb{R}^{L \times L'}$  is the mask matrix, determining whether a pair of tokens can be attended to each other. More specifically, in each Transformer block, the output vector is the concatenation of multiple attention heads  $\mathbf{H}_{\ell} = [\mathbf{A}_{\ell,1}, \dots, \mathbf{A}_{\ell,h}]$  ( $h$  is the number of heads). One attention head  $\mathbf{A}$  is computed via:

$$\mathbf{A}_{\ell} = \text{Softmax}\left(\frac{\mathbf{Q}\mathbf{K}^{\top}}{\sqrt{d_k}} + \mathbf{M}\right)\mathbf{V}, \quad (4)$$

$$\mathbf{M}_{ij} = \begin{cases} 0, & \text{allow to attend} \\ -\infty, & \text{not to attend} \end{cases} \quad (5)$$

$$\mathbf{Q} = \mathbf{W}_{\ell}^Q \mathbf{H}', \mathbf{K} = \mathbf{W}_{\ell}^K \mathbf{H}_{\ell-1}, \mathbf{V} = \mathbf{W}_{\ell}^V \mathbf{H}_{\ell-1} \quad (6)$$

where  $\mathbf{H}_{\ell-1}$  and  $\mathbf{H}'$  are linearly projected to a triple of queries, keys and values using parameter matrices  $\mathbf{W}_{\ell}^Q, \mathbf{W}_{\ell}^K, \mathbf{W}_{\ell}^V \in \mathbb{R}^{d_h \times d_k}$ , respectively;  $d_k$  is the projection dimension. In the following, we use different mask

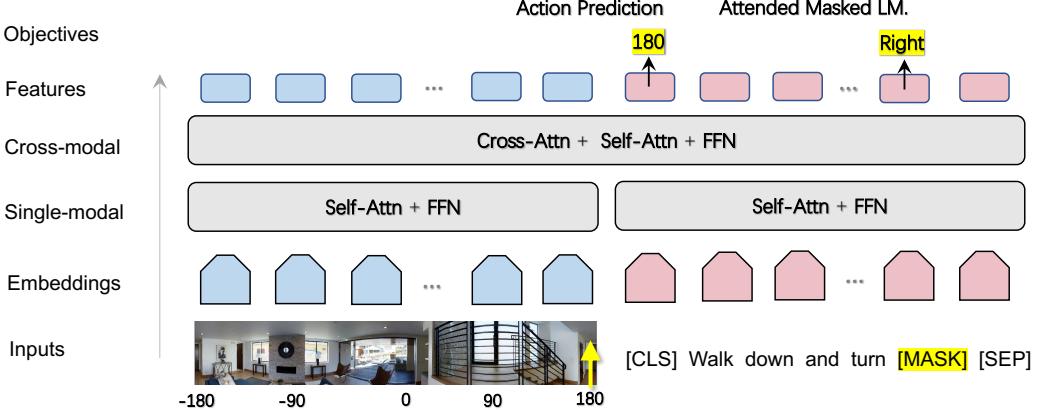


Figure 3: Illustration of the proposed pre-training model. In this example, two learning objectives are considered: (i) image-attended masked language modeling is performed on the masked word `right` in the instruction; (ii) action prediction is performed to make the decision to navigate toward direction 180. Only the language features are used for fine-tuning in downstream tasks.

matrices  $\mathbf{M}$  and attended feature matrices  $\mathbf{H}'$  to construct the contextualized representation for each module.

**Single-modal Encoder** The standard self-attention layer is used in the single-modal encoder. All of the keys, values and queries come from the output of the previous layer in the encoder. Each position in the encoder can attend to all positions that belong to its own modality in the previous layer. Specifically,  $\mathbf{M}$  is a full-zero matrix, and  $\mathbf{H}' = \mathbf{H}_{l-1}$ . Similar to the self-attention encoder module in the standard Transformer, the position-wise feed-forward network (FFN) is used.

**Cross-modal Encoder** To fuse the features from both modalities, a cross-attention layer is considered. The queries  $\mathbf{H}'$  come from the previous layer of the other modality, and the memory keys and values come from the output  $\mathbf{H}_{l-1}$  of the current modality. It allows every position in the encoder to attend over all positions in the different modality. This mimics the typical encoder-decoder attention mechanisms in the Transformer, but here we consider two different modalities, rather than input-output sequences. This cross-attention layer is followed by a self-attention layer and an FFN layer.

The overall model architecture is illustrated in Figure 3. Following [27],  $L_{\text{text}} = 9$ ,  $L_{\text{vision}} = 1$  and  $L_{\text{cross}} = 3$ . The last layer output of the encoder is denoted as  $\mathbf{z} = \mathbf{h}_{L_{\text{cross}}}$ , which is used as the features in the downstream tasks.

### 4.3. Pre-training Objectives

We introduce two main tasks to pre-train our model: Image-attended masked language modeling (MLM) and action prediction (AP). For an instruction-trajectory pair  $\{\mathbf{x}, \boldsymbol{\tau}\}$  from the training dataset  $\mathcal{D}_E$ , we assume a state-

action pair from the trajectory follows an independent identical distribution given the instruction in the pre-training stage:  $(\mathbf{s}_t, \mathbf{a}_t) \stackrel{iid}{\sim} p(\boldsymbol{\tau})$ .

**Attended Masked Language Modeling** We randomly mask out the input words with probability 15%, and replace the masked ones  $x_i$  with special token `[MASK]`. The goal is to predict these masked words based on the observation of their surrounding words  $\mathbf{x}_{\setminus i}$  and all images  $\mathbf{s}$  by minimizing the negative log-likelihood:

$$\mathcal{L}_{\text{MLM}} = -\mathbb{E}_{\mathbf{s} \sim p(\boldsymbol{\tau}), (\boldsymbol{\tau}, \mathbf{x}) \sim \mathcal{D}_E} \log p(x_i | \mathbf{x}_{\setminus i}, \mathbf{s}) \quad (7)$$

This is in analogy to the cloze task in BERT, where the masked word is recovered from surrounding words, but with additional image information to attend. It helps the learned word embeddings to be grounded in the context of visual states. This is particularly important for VLN tasks, where the agent is required to monitor the progress of completed instruction by understanding the visual images.

**Action Prediction** The output on the special token `[CLS]` indicates the fused representation of both modalities. We apply an FC layer on top of the encoder output of `[CLS]` to predict the action. It scores how well the agent can make the correct decision conditioned on the current visual image and the instruction, without referring to the trajectory history. During training, we sample a state-action pair  $(\mathbf{s}, \mathbf{a})$  from the trajectory  $\boldsymbol{\tau}$  at each step, and then apply a cross-entropy loss for optimization:

$$\mathcal{L}_{\text{AP}} = -\mathbb{E}_{(\mathbf{a}, \mathbf{s}) \sim p(\boldsymbol{\tau}), (\boldsymbol{\tau}, \mathbf{x}) \sim \mathcal{D}_E} \log p(\mathbf{a} | x_{[\text{CLS}]}, \mathbf{s}). \quad (8)$$

The full pre-training objective is:

$$\mathcal{L}_{\text{Pre-training}} = \mathcal{L}_{\text{MLM}} + \mathcal{L}_{\text{AP}}. \quad (9)$$

**Discussion** Other loss designs can be considered for the pre-training objective. Our results on masked image modeling did not show better results, and thus are excluded in the experiments.

#### 4.4. Pre-training Datasets

We construct our pre-training dataset based on the Matterport3D Simulator, a photo-realistic visual reinforcement learning (RL) simulation environment for the development of intelligent agents based on the Matterport3D dataset [5]. Specifically, it consists of two sets: (i) The training datasets of R2R, which has 104K image-text-action triplets; (ii) we employed the Speaker model in [9] to synthesize 1,020K instructions for the shortest-path trajectories on the training environments. This leads to 6,482K image-text-action triplets. Therefore, the pre-training dataset size is 6,582K.

### 5. Adapting to new tasks

We focus on three downstream VLN tasks that are based on the Matterport3D simulator. Each task poses a very different challenge to evaluate the agent. (i) The R2R task is used as an in-domain task; it can verify the agent’s generalization capability to unseen environments. (ii) CVDN and HANNA are considered as out-of-domain tasks, to study the generalization ability of the agent to new tasks. More specifically, CVDN considers indirect instructions (*i.e.*, dialog history), and HANNA is an interactive RL task.

#### 5.1. Room-to-Room

In R2R, the goal is to navigate from a starting position to a target position with the minimal trajectory length, where the target is explicitly informed via language instruction. To use the pre-trained model for fine-tuning in R2R, the attended contextualized word embeddings are fed into an LSTM encoder-decoder framework, as in [9, 16]. In prior work, random initialization is used in [9], and BERT is used in [16]. In contrast, our word embeddings are pre-trained from scratch with VLN data and tasks.

#### 5.2. Cooperative Vision-and-Dialogue Navigation

In the CVDN environment, the Navigation from Dialog History (NDH) is defined, where the agent searches an environment for a goal location, based on the dialog history that consists of multiple turns of question-answering interactions between the the agent and to its partner. The partner has privileged access to the best next steps that the agent should take according to a shortest path planner. CVDN is more challenging than R2R, in that the instructions from the dialog history are often ambiguous, under-specified, and indirect to the final target. The fine-tuning model architecture for CVDN is the same as R2R, except that CVND usually has much longer text input. We limit the sequence length to

300. Words that are longer than 300 in a dialog history are removed.

#### 5.3. HANNA: Interactive Imitation Learning

HANNA simulates a scenario where a human requester asks an agent via language to find an object in an indoor environment, without specifying the process of how to complete the task. The only source of help the agent can leverage in the environment is the *assistant*, who helps the agent by giving subtasks in the form of (i) a natural language instruction that guides the agent to a specific location, and (ii) an image of the view at that location. When the help mode is triggered, we use our pre-trained model to encode the language instructions, and the features are used for the rest of their system.

## 6. Experimental Results

#### 6.1. Training details

**Pre-training** We pre-train the proposed model on eight V100 GPUs, and the batch size for each GPU is 96. The AdamW optimizer [13] is used, and the learning rate is  $5 \times 10^{-5}$ . The total number of training epochs is 20.

**Fine-tuning** The fine-tuning is performed on NVIDIA 1080Ti GPU. For the R2R task, we follow the same learning schedule as [28]. When training the augmented listener, we use batch size 20. We continue to fine-tune the cross-attention encoder for 20k iterations, with the batch size 10 and learning rate  $2 \times 10^{-6}$ . For the NDH task, we follow the same learning schedule as in [30], and choose the batch size as 15 and learning rate as  $5 \times 10^{-4}$ . For HANNA, the training schedule is the same as [22]. The batch size is 32 and learning rate is  $1 \times 10^{-4}$ .

#### 6.2. Room-to-Room

**Dataset** The R2R dataset [3] consists of 10,800 panoramic views (each panoramic view has 36 images) and 7,189 trajectories. Each trajectory is paired with three natural language instructions. The R2R dataset consists of four splits: train, validation seen and validation unseen, test unseen. The challenge of R2R is to test the agent’s generalization ability in unseen environments.

**Evaluation Metrics** The performance of different agents is evaluated using the following metrics:

**TL Trajectory Length** measures the average length of the navigation trajectory.

**NE Navigation Error** is the mean of the shortest path distance in meters between the agent’s final location and the target location.

Agent	Validation Seen				Validation Unseen				Test Unseen				
	TL ↓	NE ↓	SR ↑	SPL ↑	TL ↓	NE ↓	SR ↑	SPL ↑	TL ↓	NE ↓	SR ↑	SPL ↑	
RANDOM	9.58	9.45	16	-	9.77	9.23	16	-	9.93	9.77	13	12	
SEQ2SEQ	11.33	6.01	39	-	8.39	7.81	22	-	8.13	7.85	20	18	
RPA	-	5.56	43	-	-	7.65	25	-	9.15	7.53	25	23	
Greedy, <b>S</b>	SPEAKER-FOLLOWER	-	3.36	66	-	-	6.62	35	-	14.82	6.62	35	28
	SMNA	-	-	-	-	-	-	-	18.04	5.67	48	35	
	RCM+SIL(TRAIN)	10.65	3.53	67	-	11.46	6.09	43	-	11.97	6.12	43	38
	REGRETFUL	-	3.23	69	63	-	5.32	50	41	13.69	5.69	48	40
	FAST	-	-	-	-	21.17	4.97	56	43	22.08	5.14	54	41
	ENVDROP	11.00	3.99	62	59	10.70	5.22	52	48	11.66	5.23	51	47
	PRESS	10.57	4.39	58	55	10.36	5.28	49	45	10.77	5.49	49	45
	PREVALENT (ours)	10.32	3.67	69	<b>65</b>	10.19	<b>4.71</b>	<b>58</b>	<b>53</b>	10.51	5.30	<b>54</b>	<b>51</b>
	<b>M</b> PRESS	10.35	3.09	71	67	10.06	4.31	59	55	10.52	4.53	57	53
	<b>M</b> PREVALENT	<b>10.31</b>	3.31	67	63	<b>9.98</b>	<b>4.12</b>	<b>60</b>	<b>57</b>	<b>10.21</b>	<b>4.52</b>	<b>59</b>	<b>56</b>
Human	-	-	-	-	-	-	-	-	11.85	1.61	86	76	

Table 1: Comparison with the state-of-the-art methods on R2R. **Blue** indicates the best value in a given setting. **S** indicates the single-instruction setting, **M** indicates the multiple-instruction setting.

**SR** Success Rate is the percentage of the agent’s final location that is less than 3 meters away from the target location.

**SPL** Success weighted by Path Length [1] trades-off SR against TL. A higher score represents more efficiency in navigation.

Among these metrics, SPL is the recommended primary metric, and other metrics are considered as auxiliary measures.

**Baselines** We compare our approach with *nine* recently published systems:

- RANDOM: an agent that randomly selects a direction and moves five step in that direction [3].
- S2S-ANDERSON: a sequence-to-sequence model using a limited discrete action space [3].
- RPA [33]: an agent that combines model-free and model-based reinforcement learning, using a lookahead module for planning.
- SPEAKER-FOLLOWER [9]: an agent trained with data augmentation from a speaker model on the panoramic action space.
- SMNA [18]: an agent trained with a visual-textual co-grounding module and a progress monitor on the panoramic action space.
- RCM+SIL [32]: an agent trained with cross-modal grounding locally and globally via RL.
- REGRETFUL [19]: an agent with a trained progress monitor heuristic for search that enables backtracking.
- FAST [12]: an agent that uses a fusion function to score and compare partial trajectories of different lengths, which enables the agent to efficiently backtrack after a mistake.

- ENVDROP [28]: an agent is trained with environment dropout, which can generate more environments based on the limited seen environments.
- PRESS [16]: an agent is trained with pre-trained language models and stochastic sampling to generalize well in the unseen environment.

**Comparison with SoTA** Table 1 compares the performance of our agent against the existing published top systems.<sup>4</sup>. Our agent PREVALENT outperforms the existing models on SR and SPL by a large margin. On both validation seen and unseen environments, PREVALENT outperforms other agents on nearly all metrics.

In PRESS [16], multiple introductions are used. To have a fair comparison, we follow [16], and report PREVALENT results. We see that testing SPL is improved. Further, the gap between seen and unseen environments of PREVALENT is smaller than PRESS, meaning that image-attended language understanding is more effective to help the agent generalize better to an unseen environment.

### 6.3. Cooperative Vision-and-Dialogue Navigation

**Dataset & Evaluation Metric** The CVDN dataset has 2050 human-human navigation dialogs, comprising over 7K navigation trajectories punctuated by question-answer exchanges, across 83 MatterPort houses [5]. The metrics for R2R can be readily used for the CVDN dataset. Further, one new metric is proposed for the NDH task:

**GP** Goal Progress measures the difference between completed distance and left distance to the goal. Larger

<sup>4</sup>The full list of leaderboard is publicly available: <https://evalai.cloudcv.org/web/challenges/challenge-page/97/leaderboard/270>

Agent	Validation Unseen			Test Unseen		
	Oracle	Navigator	Mixed	Oracle	Navigator	Mixed
RANDOM	1.09	1.09	1.09	0.83	0.83	0.83
SEQ2SEQ	1.23	1.98	2.10	1.25	2.11	2.35
PREVALENT (Ours)	<b>2.58</b>	<b>2.99</b>	<b>3.15</b>	<b>1.67</b>	<b>2.39</b>	<b>2.44</b>
SHORTEST PATH AGENT	8.36	7.99	9.58	8.06	8.48	9.76

Table 2: Results on CVDN measured by Goal Progress. Blue indicates the best value in a given setting.

Agent	SEEN-ENV				UNSEEN-ALL				
	SR ↑	SPL ↑	NE ↓	#R ↓	SR ↑	SPL ↑	NE ↓	#R ↓	
Rule	RANDOM WALK	0.54	0.33	15.38	0.0	0.46	0.23	15.34	0.0
	FORWARD 10	5.98	4.19	14.61	0.0	6.36	4.78	13.81	0.0
NO ASSISTANCE	17.21	13.76	11.48	0.0	8.10	4.23	13.22	0.0	
ANNA	88.37	63.92	1.33	2.9	47.45	25.50	7.67	5.8	
PREVALENT (Ours)	83.82	59.38	1.47	3.4	<b>52.91</b>	<b>28.72</b>	<b>5.29</b>	6.6	
Skyline	SHORTEST	100.00	100.00	0.00	0.0	100.00	100.00	0.00	0.0
	Perfect assistance	90.99	68.87	0.91	2.5	83.56	56.88	1.83	3.2

Table 3: Results on test splits of HANNA. The agent with “perfect assistance” uses the teacher navigation policy to make decisions when executing a subtask from the assistant. Blue indicates the best value.

values indicate a more efficient agent.

Three settings are considered, depending on which ground-truth action/path is employed [30]. *Oracle* indicates the shortest path, and *Navigator* indicates the path taken by the navigator. The *Mixed* supervision path means to take the navigator path if available, otherwise the shortest path. The results are in Table 2. The proposed PREVALENT significantly outperforms the Seq2Seq baseline on both validation and testing unseen environments in all settings, leading to the top position on the leaderboard<sup>5</sup>. Note that our encoder is pre-trained on R2R dataset. We observe that it can provide significant improvement when used the new task built on the CVDN dataset. This shows that the pre-trained model can adapt well on new tasks, and yields better generalization.

## 6.4. HANNA

**Dataset & Evaluation Metric** The HANNA dataset features 289 object types; the language instruction vocabulary contains 2,332 words. The numbers of locations on the shortest paths to the requested objects are restricted to be between 5 and 15. With an average edge length of 2.25 meters, the agent has to travel about 9 to 32 meters to reach its goals. Similar to R2R, SR, SPL and NE are used to evaluate the navigation. Further, one new metric is considered for this interactive task:

<sup>5</sup>The full list of leaderboard is publicly available: <https://evalai.cloudcv.org/web/challenges/challenge-page/463/leaderboard/1292>

**#R Number of requests** measures how many helps are requested by the agent.

The results are shown in Table 3. Two rule-based methods and two skyline methods are reported as references; see [22] for details. Our PREVALENT outperforms the baseline agent ANNA on the test unseen environments in terms of SR, SPL and NE, while requesting a slightly higher number of helps (#R). When measuring the performance gap between seen and unseen environments, we see that PREVALENT shows a significantly smaller difference than ANNA, e.g., (59.38-28.72=30.66) vs (63.92-25.50=38.42) for SPL. This means that the pre-trained joint representation by PREVALENT can reduce over-fitting, and generalise better to unseen environments.

## 6.5. Ablation Studies

**Is pre-training with actions helpful?** Our pre-training objective in (9) includes two losses,  $\mathcal{L}_{PA}$  and  $\mathcal{L}_{MLM}$ . To study the impact of each loss, we pre-train two model variants: one is based on the full objective  $\mathcal{L}_{PA} + \mathcal{L}_{MLM}$ , the other only uses  $\mathcal{L}_{MLM}$ . To verify its impact on new tasks, we consider CVDN first, and the results are shown in Table 4. Three types of text inputs are considered: Navigation QA, Orcale Answer, and All (a combination of both). More details are provided in the Appendix.

When  $\mathcal{L}_{PA}$  is employed in the objective, we see consistent improvement on nearly all metrics and settings. Note that our MLM is different from BERT in that the attention over images is used in the cross-layer. To verify whether the image-attended learning is necessary, we consider BERT in

Methods	Navigation QA			Oracle Answer			All		
	Oracle	Navigator	Mixed	Oracle	Navigator	Mixed	Oracle	Navigator	Mixed
$\mathcal{L}_{PA} + \mathcal{L}_{MLM}$	<b>2.80</b>	<b>3.01</b>	<b>3.28</b>	2.78	<b>3.44</b>	<b>3.38</b>	<b>2.58</b>	<b>2.99</b>	<b>3.15</b>
$\mathcal{L}_{MLM}$	2.69	3.00	3.25	2.84	3.35	3.19	2.52	2.98	3.14
BERT pre-training	2.26	2.71	2.94	2.70	2.68	3.06	2.46	2.74	2.64
BERT fine-tuning	2.39	2.03	2.51	2.23	2.41	2.52	2.32	2.93	2.28

Table 4: Ablation study of the pre-training objectives on CVDN, measured by Goal Progress. **Blue** indicates the best value.

Methods	Validation Seen				Validation Unseen				Test Unseen			
	TL ↓	NE ↓	SR ↑	SPL ↑	TL ↓	NE ↓	SR ↑	SPL ↑	TL ↓	NE ↓	SR ↑	SPL ↑
Two-stage	10.32	<b>3.67</b>	<b>0.69</b>	<b>0.66</b>	10.19	<b>4.71</b>	<b>0.58</b>	<b>0.53</b>	10.51	<b>5.30</b>	<b>0.54</b>	<b>0.51</b>
Feature-based	10.13	3.98	0.66	0.64	9.70	5.01	0.54	0.51	9.99	5.54	0.52	0.49

Table 5: Ablation study on R2R: feature-based vs fine-tuning. **Blue** indicates the better value.

two ways. (i) BERT pre-training: we apply the original MLM loss in BERT on our R2R pre-training dataset. The newly pre-trained BERT is used for fine-tuning on CVDN. (ii) BERT fine-tuning: we directly fine-tune off-the-shelf BERT on CVDN. Their performances are lower than the two variants of the proposed PREVALENT. This means our image-attended MLM is more effective for navigation tasks. More ablation studies on the pre-training objectives are conducted for HANNA, with results shown in the Appendix.

**Feature-based vs Fine-tuning** The pre-trained encoder can be used in two modes: (i) *fine-tuning* approach, where a task-specific layer is added to the pre-trained model, and all parameters are jointly updated on a downstream task. (ii) *feature-based* approach, where fixed features are extracted from the pre-trained model, and only the task-specific layer is updated. In this paper, all PREVALENT presented results generally have used the feature-based approach, as there are major computational benefits to pre-computing an expensive representation of the training data once, and then running many experiments with cheaper models on top of this representation. In the R2R dataset, we consider a *two-stage* scheme, where we fine-tune the cross-attention layers of the agent, after training via the feature-based approach. The results are reported in Table 5. We observe notable improvement with this two-stage scheme on nearly all metrics, expect the trajectory length.

**How does pre-training help generalization?** We plot the learning curves on the seen/unseen environments for R2R in Figure 4(a), and CVDN in Figure 4(b). Compared with the random initialized word embeddings in EnvDrop [28], the pre-trained word embeddings can adapt faster (especially in the early stage), and converge to higher performance in unseen environments. This is demonstrated by the SPL values in the Figure 4(a). By comparing the learning curves in Figure 4(b), we see a much smaller gap between seen and un-

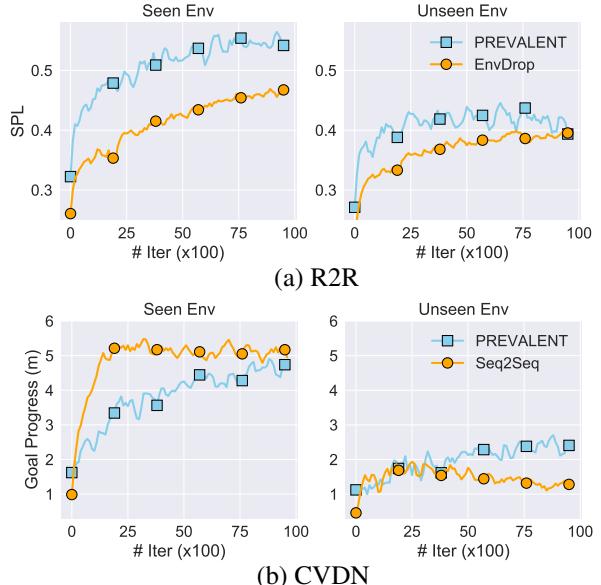


Figure 4: Learning curves on (a) R2R and (b) CVDN.

seen environments for PREVALENT than the Seq2Seq baseline [30], meaning pre-training is an effective tool to help reduce over-fitting in learning.

## 7. Conclusions

We present PREVALENT, a new pre-training and fine-tuning paradigm for vision-and-language navigation problems. This allows for more effective use of limited training data to improve generalization to previously unseen environments, and new tasks. The pre-trained encoder can be easily plugged into existing models to boost their performance. Empirical results on three benchmarks (R2R, CVDN and HANNA) demonstrate that PREVALENT significantly improves over existing methods, achieving new state-of-the-art performance.

## References

- [1] Peter Anderson, Angel Chang, Devendra Singh Chaplot, Alexey Dosovitskiy, Saurabh Gupta, Vladlen Koltun, Jana Kosecka, Jitendra Malik, Roozbeh Mottaghi, Manolis Savva, and Amir Zamir. On evaluation of embodied navigation agents. *arXiv preprint arXiv:1807.06757*, 2018. [6](#)
- [2] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. Bottom-up and top-down attention for image captioning and visual question answering. In *CVPR*, 2018. [2](#)
- [3] Peter Anderson, Qi Wu, Damien Teney, Jake Bruce, Mark Johnson, Niko Sünderhauf, Ian Reid, Stephen Gould, and Anton van den Hengel. Vision-and-language navigation: Interpreting visually-grounded navigation instructions in real environments. In *CVPR*, volume 2, 2018. [1, 3, 5, 6](#)
- [4] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016. [3](#)
- [5] Angel Chang, Angela Dai, Thomas Funkhouser, Maciej Halber, Matthias Nießner, Manolis Savva, Shuran Song, Andy Zeng, and Yinda Zhang. Matterport3D: Learning from RGB-D data in indoor environments. *International Conference on 3D Vision (3DV)*, 2017. [5, 6](#)
- [6] Howard Chen, Alane Shur, Dipendra Misra, Noah Snavely, and Yoav Artzi. Touchdown: Natural language navigation and spatial reasoning in visual street environments. *CVPR*, 2010. [1](#)
- [7] Abhishek Das, Samyak Datta, Georgia Gkioxari, Stefan Lee, Devi Parikh, and Dhruv Batra. Embodied question answering. In *CVPR*, 2018. [1](#)
- [8] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. *NAACL*, 2019. [2, 11](#)
- [9] Daniel Fried, Ronghang Hu, Volkan Cirik, Anna Rohrbach, Jacob Andreas, Louis-Philippe Morency, Taylor Berg-Kirkpatrick, Kate Saenko, Dan Klein, and Trevor Darrell. Speaker-follower models for vision-and-language navigation. *NIPS*, 2018. [1, 2, 3, 5, 6, 11](#)
- [10] Ross Girshick. Fast R-CNN. In *CVPR*, 2015. [2](#)
- [11] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. [3](#)
- [12] Liyiming Ke, Xiujun Li, Yonatan Bisk, Ari Holtzman, Zhe Gan, Jingjing Liu, Jianfeng Gao, Yejin Choi, and Siddhartha Srinivasa. Tactical rewind: Self-correction via backtracking in vision-and-language navigation. *CVPR*, 2019. [2, 6](#)
- [13] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. [5](#)
- [14] Eric Kolve, Roozbeh Mottaghi, Daniel Gordon, Yuke Zhu, Abhinav Gupta, and Ali Farhadi. AI2-THOR: An interactive 3D environment for visual AI. *arXiv preprint arXiv:1712.05474*, 2017. [1](#)
- [15] Gen Li, Nan Duan, Yuejian Fang, Dixin Jiang, and Ming Zhou. Unicoder-VL: A universal encoder for vision and language by cross-modal pre-training. *arXiv preprint arXiv:1908.06066*, 2019. [2](#)
- [16] Xiujun Li, Chunyuan Li, Qiaolin Xia, Yonatan Bisk, Asli Celikyilmaz, Jianfeng Gao, Noah Smith, and Yejin Choi. Robust navigation with language pretraining and stochastic sampling. *EMNLP*, 2019. [2, 5, 6](#)
- [17] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. ViLBERT: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *NIPS*, 2019. [2](#)
- [18] Chih-Yao Ma, Jiasen Lu, Zuxuan Wu, Ghassan AlRegib, Zsolt Kira, Richard Socher, and Caiming Xiong. Self-monitoring navigation agent via auxiliary progress estimation. *ICLR*, 2019. [1, 2, 6](#)
- [19] Chih-Yao Ma, Zuxuan Wu, Ghassan AlRegib, Caiming Xiong, and Zsolt Kira. The regretful agent: Heuristic-aided navigation through progress estimation. *CVPR*, 2019. [2, 6](#)
- [20] Dipendra Misra, John Langford, and Yoav Artzi. Mapping instructions and visual observations to actions with reinforcement learning. *EMNLP*, 2017. [1](#)
- [21] Will Monroe, Robert XD Hawkins, Noah D Goodman, and Christopher Potts. Colors in context: A pragmatic neural model for grounded language understanding. *TACL*, 2017. [1](#)
- [22] Khanh Nguyen and Hal Daumé III. Help, anna! visual navigation with natural multimodal assistance via retrospective curiosity-encouraging imitation learning. *EMNLP*, 2019. [1, 5, 7](#)
- [23] Manolis Savva, Angel X Chang, Alexey Dosovitskiy, Thomas Funkhouser, and Vladlen Koltun. MINOS: Multi-modal indoor simulator for navigation in complex environments. *arXiv preprint arXiv:1712.03931*, 2017. [1](#)
- [24] Weijie Su, Xizhou Zhu, Yue Cao, Bin Li, Lewei Lu, Furu Wei, and Jifeng Dai. VL-BERT: Pre-training of generic visual-linguistic representations. *arXiv preprint arXiv:1908.08530*, 2019. [2](#)
- [25] Chen Sun, Austin Myers, Carl Vondrick, Kevin Murphy, and Cordelia Schmid. VideoBERT: A joint model for video and language representation learning. *ICCV*, 2019. [2](#)
- [26] Ilya Sutskever, Oriol Vinyals, and Quoc V Le. Sequence to sequence learning with neural networks. In *NIPS*, 2014. [1](#)
- [27] Hao Tan and Mohit Bansal. LXMERT: Learning cross-modality encoder representations from transformers. *EMNLP*, 2019. [2, 4](#)
- [28] Hao Tan, Licheng Yu, and Mohit Bansal. Learning to navigate unseen environments: Back translation with environmental dropout. *EMNLP*, 2019. [2, 5, 6, 8](#)
- [29] Jesse Thomason, Daniel Gordon, and Yonatan Bisk. Shifting the baseline: Single modality performance on visual navigation & qa. In *NAACL*, 2019. [3](#)
- [30] Jesse Thomason, Michael Murray, Maya Cakmak, and Luke Zettlemoyer. Vision-and-dialog navigation. *CoRL*, 2019. [1, 5, 7, 8](#)
- [31] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NIPS*, 2017. [3](#)
- [32] Xin Wang, Qiuyuan Huang, Asli Celikyilmaz, Jianfeng Gao, Dinghan Shen, Yuan-Fang Wang, William Yang Wang, and Lei Zhang. Reinforced cross-modal matching and self-supervised imitation learning for vision-language navigation. *CVPR*, 2019. [1, 2, 6](#)

- [33] Xin Wang, Wenhan Xiong, Hongmin Wang, and William Yang Wang. Look before you leap: Bridging model-free and model-based reinforcement learning for planned-ahead vision-and-language navigation. *ECCV*, 2018. [1](#), [2](#), [6](#)
- [34] Luowei Zhou, Hamid Palangi, Lei Zhang, Houdong Hu, Jason J Corso, and Jianfeng Gao. Unified vision-language pre-training for image captioning and VQA. *AAAI*, 2020. [2](#), [11](#)

## Supplementary Material: Towards Learning a Generic Agent for Vision-and-Language Navigation via Pre-training

**Summary of Contributions.** Weituo implemented the algorithm, made the model work, and ran all experiments. Chunyuan initiated the idea of pre-training the first generic agent for VLN, led and completed the manuscript writing. Xiujun provided the codebase and helped implementation. Lawrence and Jianfeng edited the final manuscript.

### A. Pre-training Dataset Preparation

We found that the largest VLN training dataset R2R contains only 104K samples, an order magnitude smaller than the pre-training datasets typically used in language [8] or vision-and-language pre-training [34]. This renders a case where pre-training can be degraded due to insufficient training data, while harvesting such samples with human annotations is expensive. Fortunately, we can resort to generative models to synthesize the samples. We first train an seq2seq auto-regressive model (*i.e.*, a speaker model [9]) that can produce language instructions conditioned on the agent trajectory (a sequence of actions and visual images) on R2R dataset; then collect a large number of shortest trajectories using the Matterport 3D Simulator, and synthesize their corresponding instructions using the speaker model. This leads to 6482K new training samples. The two datasets are compared in Figure 4(b). The agent is pre-trained on the combined dataset. Our results show that synthetic samples produced by generative models can be incorporated into the pre-training data and helps self-supervised learning.

### B. Experiments

**Three types of inputs on CVDN** We illustrate the naming of three types of text inputs on CVDN in Table 6.

	$V$	$t_0$	$A_i$	$Q_i$	$Q_{1:i-1} \& A_{1:i-1}$
Oracle Answer	✓	✓	✓		
Navigation QA	✓	✓	✓	✓	
All	✓	✓	✓	✓	✓

Table 6: Three types of inputs on CVDN.  $t_0$  is the target object,  $V$  is the ResNet feature.  $Q_i$  and  $A_i$  are the question and answers in the  $i$ -th turn.  $Q_{1:i-1} \& A_{1:i-1}$  are the question & answer pairs before the  $i$ -th turn.

**Ablation Study Results on HANNA** Table 7 shows the results with different pre-training objectives. We see that the  $\mathcal{L}_{PA} + \mathcal{L}_{MLM}$  yields the best performance among all variants.

Agent	SEEN-ENV				UNSEEN-ALL			
	SR $\uparrow$	SPL $\uparrow$	NE $\downarrow$	#R $\downarrow$	SR $\uparrow$	SPL $\uparrow$	NE $\downarrow$	#R $\downarrow$
PREVALENT ( $\mathcal{L}_{PA} + \mathcal{L}_{MLM}$ )	<b>83.82</b>	<b>59.38</b>	<b>1.47</b>	<b>3.4</b>	<b>52.91</b>	<b>28.72</b>	<b>5.29</b>	<b>6.6</b>
PREVALENT ( $\mathcal{L}_{MLM}$ )	78.75	54.68	1.82	4.3	44.29	24.27	6.33	8.1
BERT (feature-based)	57.54	34.33	4.71	3.9	24.12	11.50	9.55	11.3
BERT (fine-tuning)	80.75	57.46	1.97	4.0	26.36	12.66	9.1	8.3

Table 7: Ablation study of pre-training objectives on test splits of HANNA.



Figure 5: The percentage of pre-training datasets. The synthesized dataset occupies 98.4%.

## C. Comparison with Related Work

**Comparison with PRESS.** The differences are summarized in Table 8 (a). Empirically, we show that (1) incorporating visual and action information into pre-training can improve navigation performance; (2) Pre-training can generalize across different new navigation tasks.

**Comparison with vision-language pre-training (VLP).** The differences are in Table 8 (b). Though the proposed methodology generally follows self supervised learning such as VLP or BERT, our research scope and problem setups are different, which renders existing pre-models are not readily applicable.

	<b>Prevalent</b> (Proposed)	<b>Press</b>
<b>Dataset</b>	Augmented R2R dataset	Generic language
<b>Modality</b>	Vision-language-action triplets	Language
<b>Learning</b>	Train from scratch	Off-the-shelf (BERT)
<b>Downstream</b>	Three navigation tasks	R2R

	<b>Prevalent</b> (Proposed)	<b>VLP</b>
<b>Visual Input</b>	Panoramic views (Size: $36 \times 640 \times 480$ )	Single image (Size: $640 \times 480$ )
<b>Visual Features</b>	ResNet (View-level)	Fast RCNN (Object-level)
<b>Objectives</b>	Attentive MLM & Action Prediction	Masking on VL & Same-Pair Prediction
<b>Downstream</b>	RL: Navigation in sequential decision-making environments	Single-step prediction

(a) PRESS

Table 8: Comparison with related works.

(b) VLP