Learning Disentanglement with Decoupled Labels for Vision-Language Navigation

Wenhao Cheng[1]†[0000−0003−4108−1647], Xingping Dong[2]†[0000−0003−1613−9288], Salman Khan[3], and Jianbing Shen[4]⋆[0000−0003−1883−2086]

[1] School of Computer Science, Beijing Institute of Technology
[2] Inception Institute of Artificial Intelligence, UAE
[3] Mohamed bin Zayed University of Artificial Intelligence, UAE
[4] SKL-IOTSC, Computer and Information Science, University of Macau

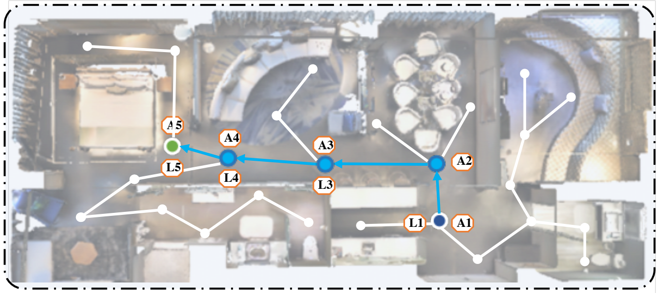
Abstract. Vision-and-Language Navigation (VLN) requires an agent to follow complex natural language instructions and perceive the visual environment for real-world navigation. Intuitively, we find that instruction disentanglement for each viewpoint along the agent's path is critical for accurate navigation. However, most methods only utilize the whole complex instruction or inaccurate sub-instructions due to the lack of accurate disentanglement as an intermediate supervision stage. To address this problem, we propose a new *Disentanglement framework with Decoupled Labels* (DDL) for VLN. Firstly, we manually extend the benchmark dataset Room-to-Room with landmark- and action-aware labels in order to provide fine-grained information for each viewpoint. Furthermore, to enhance the generalization ability, we propose a Decoupled Label Speaker module to generate pseudo-labels for augmented data and reinforcement training. To fully use the proposed fine-grained labels, we design a Disentangled Decoding Module to guide discriminative feature extraction and help alignment of multi-modalities. To reveal the generality of our proposed method, we apply it on a LSTM-based model and two recent Transformer-based models. Extensive experiments on two VLN benchmarks (i.e., R2R and R4R) demonstrate the effectiveness of our approach, achieving better performance than previous state-of-the-art methods.

Keywords: Vision-and-Language Navigation, Disentanglement, Modular Network, Imitation/Reinforcement learning, LSTM and Transformer

1 Introduction

Vision-and-language navigation is a challenging task that requires the agent to perceive its visual environment and understand the natural language instructions to reach the target location. Recent works have achieved remarkable progress via techniques such as pre-exploration [74,33,47,71,13], pre-training [23,48,40,41,45,21], reward shaping [72,56,73], auxiliary tasks [79,46,76], data augmentation [18,65,32] and counterfactual thinking [19,53,69].

⋆ Corresponding author: *Jianbing Shen* (shenjianbingcg@gmail.com). † Equal contribution. Codes and annotations are available at https://github.com/cwhao98/DDL.



Instruction:
Go straight$^{\text{A1}}$ to the white chairs$^{\text{L1}}$. Turn left and work forward$^{\text{A2}}$. Pass$^{\text{A3}}$ the couches on the right$^{\text{L3}}$ and go into$^{\text{A4}}$ the room straight ahead$^{\text{L4}}$. Wait$^{\text{A5}}$ by the bed$^{\text{L5}}$.

Fig. 1. An illustration of decoupled labels providing intermediate supervision during navigation. The superscripts in the instruction denote the landmark and action labels for each viewpoint. The decoupled labels not only contain disentangled information, but help the alignment between vision and language modalities.

The existing VLN dataset Room-to-Room [2] only provides complex human instructions which contain information about several different attributes, *e.g.*, objects, landmarks and actions. Such convoluted instructions make the agent's task more challenging. Our intuition is that disentangling these instructions can provide more accurate and clear input to improve the decisions taken by the agent. This idea is also inspired by the human concept [6], since human beings usually do orthogonal decomposition for cognition, *i.e.*, divide something into different attributes to better understand and remember. In particular, the previous works OAAM [56] and RelGraph [26] tried to disentangle the instructions into different kinds of information via attention mechanism [68]. However, the attention-based models can produce inaccurate disentangled instructions which can mislead the agent, resulting in a performance degradation.

In addition, the alignment between vision and language is also a challenging open issue in VLN. To alleviate this issue, RxR [38] provides time-aligned multilingual instruction, but without the decoupling of specific parts such as landmarks and actions. FGR2R [27] and BabyWalk [80] split the long instruction into small parts via chunking function and dynamic programming, since sub-instructions are more conducive to match visual scenes. Specifically, FGR2R utilizes a shifting module to predict the alignment between sub-instruction and navigation path. However, such a split is not fine-grained enough to provide accurate decoupled labels to achieve proper disentanglement in the VLN task.

To address the above issues and provide fine-grained guidance for VLN, we propose a novel *Disentanglement framework with Decoupled Labels* (DDL). Our framework has three main highlights: fine-grained labels, a decoupled label speaker, and a disentangled decoding module, which we elaborate below.

Fine-Grained Labels. We enrich the benchmark R2R [2] by adding new fine-grained human annotations, and call it Landmark- and Action-aware Room-to-Room Labels (LAR2R). Specifically, as shown in Fig. 1, for each viewpoint, we annotate the specific landmark and action sub-instructions that should be highlighted to navigate correctly at the current viewpoint. Therefore, the annotated labels not only provide more precise disentangled instruction, but implicitly contain the alignment information between multi-modalities.

Decoupled Label Speaker. Most recent VLN models are trained by both Imitation Learning (IL) and Reinforcement Learning (RL), while we can only use the annotated labels during IL training since the paths in RL training phase are unknown and abundant, and it is impossible to annotate so many paths. Moreover, augmentation methods are often adopted, such as back translation augmentation [65] and random environmental mixup [44], to obtain more trajectories for training. These trajectories are also not annotated. To provide supervised signal for RL training phase and augmented data, we propose a decoupled label speaker. Taking the given instruction and visual observation along a trajectory as input, our speaker module can generate landmark and action pseudo-labels, enabling most VLN models to be trained with decoupled labels.

Disentangled Decoding Module. To make full use of the proposed fine-grained labels, we design a Disentangled Decoding Module to guide discriminative feature extraction and help the alignment of multiple modalities. Specifically, given a VLN model, we firstly design a disentanglement branch, based on its feature encoding backbone, to enable decoupling. Then, we employ a language auxiliary loss that uses the decoupled labels to regularize the landmark- and action-aware attention weights, making complex inputs easier to understand for the agent. Note that our approach is *model-agnostic* and can easily be integrated into most VLN methods. We adopt three representative algorithms: a LSTM-based navigator OAAM [56], two Transformer-based navigators VLN↻BERT [28] and HAMT [11], as baselines to show the generality of our proposed approach.

Our main contributions are summarized as follows: 1) We develop a new Disentanglement framework with Decoupled Labels (DDL) for the VLN task. DDL uses decoupled labels to guide the extraction of disentangled features and help the alignment between vision and language modalities, making the navigation more interpretable. 2) We enrich the benchmark dataset R2R [2] with landmark- and action-aware annotations. To the best of our knowledge, this is the first effort to demonstrate the effectiveness of fine-grained decoupled labels in VLN. 3) To enhance generalization ability, we further propose a decoupled label speaker to generate pseudo-labels for reinforced learning and augmented data. In addition, our speaker can be easily integrated into most VLN models to provide fine-grained labels. 4) To reveal the generality of our DDL, we apply it to both LSTM-based and Transformer-based methods. Extensive experiments on R2R [2] and R4R [32] demonstrate the improvement over three competitive baselines and state-of-the-art performance of our models.

2 Related Work

Vision-and-Language Navigation. Recently VLN has attracted significant research interest. Supported by various simulators [7,35,62], a number of tasks such as R2R [2], REVERIE [57], ALFRED [64], CVDN [67], HANNA [51], and VNLA [52] have been proposed. Many early approaches [18,46,46,47,33] for R2R are based on Imitation Learning (IL), since the agent can learn quickly from teacher actions through Behaviour Cloning [5]. Speaker-Follower [18] introduces a speaker to synthesize new instructions. Self-Monitoring [46] proposes a progress monitor for VLN agent. In addition to IL, RL-based methods have also achieved great success with strong generalization ability. RPA [74] first combines model-free and model-based deep RL for navigation. RCM [73] enforces cross-modal grounding both locally and globally. E-Drop [65] uses the environmental dropout method to generate more unseen environments. Other approaches try to improve performance by auxiliary tasks [79], reward shaping via distillation [72], active perception [71], structured scene memory [70], 3D semantic representation [66], contrastive learning [42], snapshot ensemble [59] and counterfactual cycle-consistent [69]. Due to the success of transformer [68] and BERT [14], many transformer architectures [17,40,39,10,48,12] for VLN have emerged. VLN⟳BERT [28] introduces a recurrent unit within transformer to enable past information flow. HAMT [11] and E.T. [54] encode all the observation and action history within a full transformer. MTVM [43] proposes variable-length memory to encode history information. Concurrently, HOP [58] designs proxy tasks to model spatio-temporal alignment, further mining the role of historical information. SEvol [9] constructs object-level layout graph to maintain navigation state with a reinforced state evolving strategy. Apart from the above approaches that focus mainly on indoor navigation, VLN in outdoor scenes [8], continuous environments [37,36,60,31] and multilingual navigation with spatial-temporal grounding [38] have also been explored.

Disentangled Representation in VLN. Intuitively, disentangling the instruction or visual scene will help the agent better understand the complex input. Early work [29] has explored the effectiveness of grounding language to multiple modalities. Recently, OAAM [56] utilizes two learning attention modules to disentangle the object- and action-related parts in the instruction. Hong *et al.* [26] build a language and visual graph to capture the relationship of scenes, objects, and direction clues. ORIST [55] leverages object- and word-level feature representations to facilitate modality matching. CKR [20] decouples the room-type and object-entity explicitly, incorporating knowledge graph to help the entity reasoning. SOAT [50] encodes the scene feature and object reference separately in transformer which leads to performance improvement. However, the above methods are often based on attention mechanisms, which can generate inaccurate results. Although multi-head self-attention within transformers attends to information from different subspaces, it is not easily interpretable. Therefore, in this paper, we investigate the effect of decoupled labels to guide discriminative feature extraction, making VLN better interpretable.

3 Our Approach

In this section, we first formulate the VLN task in 3.1 and then briefly summarize the three baseline navigators, OAAM [56], VLN⟳BERT [28] and HAMT [11], in 3.2. The Landmark- and Action-aware Room-to-Room (LAR2R) labels introduced in our work are explained in 3.3. The proposed disentanglement framework is outlined in 3.4. We further explain how to get pseudo-labels for unlabeled data in 3.5. Finally, we present the model training details in 3.6.

3.1 Problem Setup

The standard VLN task requires the agent to navigate in a connected graph to the target location following natural language instruction. Formally, given an instruction I of L words, $I = \{w_1, w_2, \ldots, w_L\}$, at each time step t, the agent obtains the surrounding environment information, which is discretized into 36 single view images $\{v_{t,i}\}_{i=1}^{36}$. Each view $v_{t,i}$ is represented by visual feature $f_{t,i}$ and orientation feature $o_{t,i} = (\cos\theta_{t,i}, \sin\theta_{t,i}, \cos\phi_{t,i}, \sin\phi_{t,i})$, where $v_{t,i}$ is an image at orientation with heading angle $\theta_{t,i}$ and elevation angle $\phi_{t,i}$. The image feature $f_{t,i}$ can be obtained by a detector [1], or ResNet [24] pretrained on Imagenet [61] or Place365 [78]. Besides, there are N_t candidate directions for the agent to select at each viewpoint, where the set of view features for each candidate direction is given by $\{c_{t,k}\}_{k=1}^{N_t}$ which are of the same type as $v_{t,i}$.

3.2 Conventional Navigation

To showcase the generality of porposed approach, we experiment on three recent navigators, OAAM [56], VLN↻BERT [28] and HAMT [11]. OAAM is a LSTM-based navigator while VLN↻BERT and HAMT [11] are transformer-based navigators. All three take natural language instruction and visual perception as input, and output the selected actions across several candidate directions at each step. However, their architectures are very different partly in language encoding, decision making, and the maintenance of internal state during navigation.

LSTM-Based Navigator. OAAM [56] is built upon EnvDrop [65], which firstly encodes the language instruction by a Bi-LSTM at the beginning of navigation, and then utilizes another LSTM to enable the entire navigation process. At each step t, the agent updates its internal state h_t by previous latent state and instruction-aware visual observation at the current viewpoint. Formally:

$$h_t = \mathrm{LSTM}(h_{t-1}, [o_t, I]) \tag{1}$$

where I is the instruction encoding, o_t is the perceived panoramic view feature, and $[\cdot]$ denotes concatenation.

In terms of the navigation decision making, OAAM adopts two learnable attention modules to highlight the corresponding object- and action-related part of the given instruction which are fed into the object-vision and action-orientation matching modules, respectively, to predict the selected direction. This is followed by an adaptive module to combine the action logits as the final decision. For more details, please refer to the supplementary materials.

Transformer-Based Navigator. VLN↻BERT [28] is a state-of-the-art agent that introduces a recurrent unit to the transformer, which enables information flow from the past to the current state during the entire navigation process. In

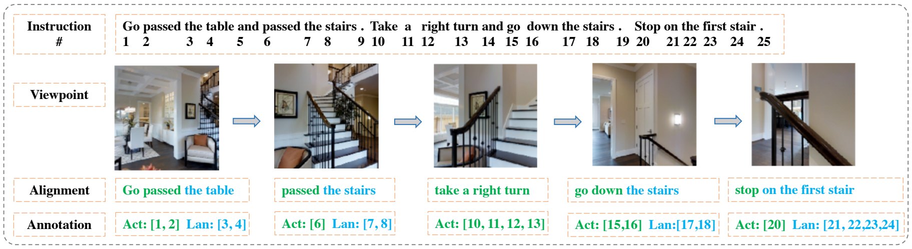


Fig. 2. Navigation with specific landmark- and action-aware sub-instructions in LAR2R. We extend the R2R dataset by providing annotations of landmark (blue) and action (green) related parts in the instruction along with each viewpoint.

contrast, HAMT [11] processes all historical information to enrich the current representation. The language instruction is encoded via multi-head self-attention at the beginning and the leading input token [CLS] is selected as the agent's initial state. Then, at each time step, the agent takes the language tokens and observed visual features as input, which are processed via cross-modal attention and then followed by self-attention on each candidate view to update the internal state and visual tokens. Formally, this is represented as:

$$h_t = \mathrm{BERT}([\mathrm{CLS}], I, o_t, p_t) \tag{2}$$

where h_t is the current agent state, [CLS] is a pre-defined classification token in the BERT model, and p_t denotes past history input that is only used in HAMT. To take the decision on next direction, attention scores over each candidate will be used as the action probability by the agent.

3.3 The LAR2R Labels

Label Collection. In order to better decouple the information of different attributes in the input instruction and to help the agent locate the specific sub-instruction part, we extend the R2R dataset with fine-grained annotations. Specifically, as shown in Fig. 2, at each navigation step t, we annotate the landmark part L_t and action part A_t in the instruction that should be attended to select the next action at the current viewpoint. Formally, we have:

$$L_t = [l_{t,1}, \ldots, l_{t,N_l}], \quad A_t = [a_{t,1}, \ldots, a_{t,N_a}], \tag{3}$$

where $l_{t,i}$ and $a_{t,j}$ are the index of landmark- and action-related instructions respectively. N_l and N_a are the total number of words that should be highlighted at the current time step. To maintain consistency and ensure accuracy, we ask one of the annotators to mark the labels which is crosschecked by another person. The overall process took about four months of annotation effort.

Label Statistics. For the training split, we have annotated 40,813 viewpoints for landmark-related instruction and 52,735 viewpoints for action-related instruction, with 3.6 and 1.9 words on average for each viewpoint, respectively.

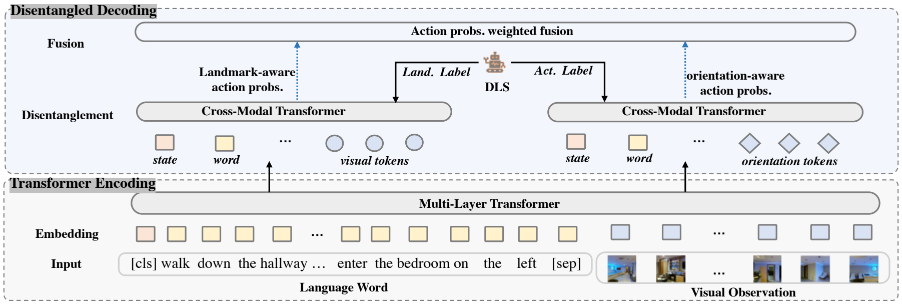


Fig. 3. Overview of the Transformer-based Disentangled Decoding Module. The model takes language words and visual observation as input. After the transformer encoding, two parallel cross-modal transformers are utilized to enable disentangled decoding, supervised by our decoupled labels via a language auxiliary loss. Then the output of two disentangled branches is fused to predict the final action of the agent. DLS represents the decoupled label speaker (see Sec. 3.5 for details).

Note that some sub-instructions will cover two or more viewpoints (*e.g.*, go up the stairs all the way), which we only annotate once at the first location. For the validation set, the landmark-points and action-points are 6,841 and 8,732 for validation unseen, 3,058 and 3,946 for validation seen split, respectively. In total, LAR2R provides about 1,15,000 image-text pairs.

3.4 Disentangled Decoding Module

Our method aims to use accurate labels to guide the disentangled feature extraction. Therefore, a prerequisite is that the model architecture allows feature disentanglement. Given a model without decoupling, we first propose a simple way to achieve disentanglement, and then boost the performance using accurate annotated labels and pseudo-labels with a language auxiliary loss (LAL). Since the original architecture of OAAM [56] uses separate streams to process object and action related cues, our approach is feasible to be directly integrated with it. We use the proposed LAL to optimize the attention weight of two learnable attention modules in parallel streams within OAAM. Next, we take Transformer-based methods as an example to illustrate the proposed approach. The overview is presented in Fig. 3. Note that the history input in HAMT is omitted for brevity. We explain the module architecture below.

Transformer Block Notation. Each Transformer block encodes features from previous block X_{l-1}, consisting of Multi-Head Self Attention (MSA) and Multi-Layer Perception (MLP) with residual connections and layer normalization. Formally, we can denote one Transformer block as:

$$\begin{aligned} H_l &= \mathrm{LN}(\mathrm{MSA}(X_{l-1}) + X_{l-1}) \\ Z_l &= \mathrm{MLP}(H_l) \\ X_l &= \mathrm{LN}(Z_l W_l + H_l) \end{aligned} \tag{4}$$

where LN is layer normalization [4], W_l is a learnable projection matrix, and Z_l is an intermediate output that increases the feature dimension of H_l through MLP to obtain more powerful representations.

As the term suggests, MSA in Eq. 4 captures dependencies between the tokens obtained from the input sequence elements using scaled dot-product attention ($Attn$). For MSA, the queries, keys, and values are generated from the same input i.e., $\mathrm{MSA}(X) = Attn(W_q X,\ W_k X,\ W_v X)$ using learned projection matrices W_q, W_k, W_v. To enable Multi-head Cross-modal Attention (MCA), we denote $\mathrm{MCA}(U,\ V) = Attn(W_q V,\ W_k U,\ W_v U)$. MCA uses the features V in one modality to query their correlation with the features U of another modality.

Disentanglement Branch. At each navigation step t, the agent observes k candidate directions where each view i is composed of visual feature $f_{t,i} \in \mathbb{R}^{d1}$ and orientation feature $o_{t,i} \in \mathbb{R}^{d2}$. To disentangle the observation, we design two BERT blocks to process the visual and geometry clues separately. Firstly, we have:

$$\hat{f}_{t,i} = f_{t,i} W_f \quad \hat{o}_{t,i} = o_{t,i} W_o, \tag{5}$$

where $W_f \in \mathbb{R}^{d1 \times d}$ and $W_o \in \mathbb{R}^{d2 \times d}$ are learnable parameters to project the features into the same space as the language tokens. Meanwhile, we encode the agent state $h_t \in \mathbb{R}^d$ to get a transformed representation $\hat{h}_t$ via:

$$\hat{h}_t = \mathrm{Tanh}(h_t W_h). \tag{6}$$

Next, to highlight the landmark- and action-aware instruction, the refined state $\hat{h}_t \in \mathbb{R}^d$ will be concatenated with the two types of disentangled tokens respectively, and fed into the cross-modal attention block. Formally:

$$\mathrm{E}_{lan} = \mathrm{LN}(\mathrm{MCA}(\mathrm{C}_{lan}, I) + \mathrm{C}_{lan}), \tag{7}$$

$$\mathrm{E}_{act} = \mathrm{LN}(\mathrm{MCA}(\mathrm{C}_{act}, I) + \mathrm{C}_{act}), \tag{8}$$

where I is the encoded language instruction, $\mathrm{C}_{lan} = [\hat{h}_t, \hat{f}_{t,1}, ..., \hat{f}_{t,k}] \in \mathbb{R}^{(k+1) \times d}$ and $\mathrm{C}_{act} = [\hat{h}_t, \hat{o}_{t,1}, ..., \hat{o}_{t,k}] \in \mathbb{R}^{(k+1) \times d}$. Note that past history information will also be included here for the case of HAMT.

Subsequently, to get the intermediate action probability for each visual direction, multi-head cross attention will be performed on E_{lan} and E_{act}. The landmark-aware score $\mathrm{A}_{lan,t} \in \mathbb{R}^k$ and action-aware score $\mathrm{A}_{act,t} \in \mathbb{R}^k$ will be calculated by the average attention weight of all heads over each candidate token $\hat{f}_{t,i}$ and $\hat{o}_{t,i}$, respectively. At last, we perform a fusion operation where the final action probability $P_t \in \mathbb{R}^k$ for each candidate is the weighted sum over the output of two disentanglement branches:

$$\mathrm{P}_t = \mathrm{Softmax}([\mathrm{A}_{lan,t}, \mathrm{A}_{act,t}] W_s), \tag{9}$$

where $W_s = \hat{h}_t W_x$, and $W_x \in \mathbb{R}^{d \times 2}$, $W_s \in \mathbb{R}^2$ are learnable parameters, which decide the attended language component at the current position.

Language Auxiliary Loss. Given the index of landmark- and action-aware instruction, an intuitive idea is to utilize the label to regularize the attended language attention weight. Thus, we propose a language auxiliary loss (LAL) to guide more accurate disentangled feature extraction.

Considering the cross-modal attention block (Eqs. (7) and (8)), both the state token ($\hat{h}_t$) and candidate tokens ($\hat{f}_{t,i}$ and $\hat{o}_{t,i}$) will attend to the language instruction. Instead of regularizing all attention weights of each token, we only optimize those of the state token, since the later self-attention will send the disentangled information to each candidate. Specifically, the landmark-aware language attention weight $\bar{\gamma}^n_{t,j} \in \mathbb{R}^1$ at step t is formulated by:

$$\bar{\gamma}^n_{t,j} = \frac{Q^n_t {K^n_{t,j}}^\intercal}{\sqrt{d_h}}, \tag{10}$$

where d_h is the dimension of hidden state, j represents the index of a word in the instruction, n denotes the index of attention head, and Q is the query of agent state $\hat{h}_t$ while K is the key generated by each textual token. Then, to deal with the case where some states attend to more than one instruction word, a Sigmoid function is applied on the average attention weight of each head:

$$\gamma_{t,j} = \text{Sigmoid}(\frac{1}{N}\sum_{n=1}^{N} \bar{\gamma}^n_{t,j}). \tag{11}$$

Similarly, the action-aware language attention weight $\sigma_{t,j} \in \mathbb{R}^1$ can be obtained. Finally, a Binary Cross Entropy loss is enforced, as follows:

$$\mathcal{L}_{lan} = -\frac{1}{TL}\sum_{t=1}^{T}\sum_{j=1}^{L} x_{t,j}\log(\gamma_{t,j}) + (1 - x_{t,j})\log(1 - \gamma_{t,j}), \tag{12}$$

$$\mathcal{L}_{act} = -\frac{1}{TL}\sum_{t=1}^{T}\sum_{j=1}^{L} y_{t,j}\log(\sigma_{t,j}) + (1 - y_{t,j})\log(1 - \sigma_{t,j}), \tag{13}$$

where T is the number of navigation steps and L is instruction length, $\gamma_{t,j}$ and $\sigma_{t,j}$ are the predicted attention weights of landmark and action at each time step t. $x_{t,j}$ and $y_{t,j}$ are binary labels which are assigned to 1 only when the j-th word index is in L_t and A_t in Eq. (3), respectively.

3.5 Decoupled Label Speaker

Incorporating the baseline model and the Disentangled Decoding Module, we can utilize the proposed fine-grained labels to enhance the discriminative ability of the agent. However, our fine-grained label is only effective during Imitation Learning (IL) with original training data, since the exploration in Reinforcement Learning (RL) will produce abundant trajectories without fine-grained labels. Moreover, most VLN approaches use augmentation methods [65,44] to obtain more trajectories for training. These trajectories are also not annotated and can not be used by our model. A common idea is to use the baseline model

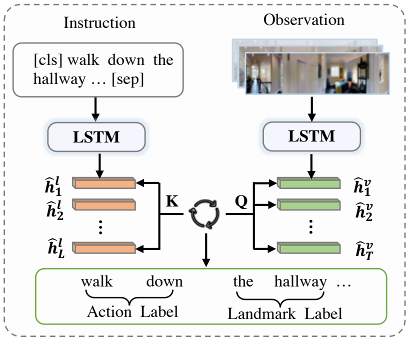


Fig. 4. Architecture of the proposed Decoupled Label Speaker (DLS). Taking the language instruction and visual observation as inputs, the DLS first encodes them using LSTMs and then employs cross-modal attention to predict landmark and action labels for each viewpoint.

to generate pseudo-labels, but this is only a by-product during the navigation. Moreover, the baseline is too large to be applied to other models. Therefore, we propose a general Decoupled Label Speaker (DLS) to provide supervised signals with landmark and action labels for these generated trajectories.

In general, as shown in Fig. 4, the DLS adopts an encoder-decoder paradigm. We first utilize two LSTM [25] to encode the observations along the path and corresponding instructions respectively, and then two cross-modal attention modules are imposed as landmark- and action-speaker to disentangle the instruction for each viewpoint.

Specifically, we use an external memory to store the visual observation during navigation, and then a Bi-LSTM is used to capture context information:

$$[h_1, \ldots, h_T] = \text{Bi-LSTM}(c_1, \ldots, c_T), \tag{14}$$

where T is the length of trajectory, and c_i is the view feature of selected candidate direction. Then, we attend the panoramic view o_t with the hidden state h_t:

$$\begin{aligned} z_{t,i} &= \text{Softmax}_i(o_{t,i}^T W_z h_t), \\ h_t^v &= \sum_i z_{t,i} o_{t,i}, \\ \hat{h}_t^v &= \text{Tanh}(W_v[h_t; h_t^v]), \end{aligned} \tag{15}$$

where $\hat{h}_t^v$ is the vision-aware hidden state at each viewpoint, and W_v and W_z are trainable parameters. To get the instruction-aware hidden state, we use another encoder:

$$[\hat{h}_1^l, \ldots, \hat{h}_L^l] = \text{LSTM}(\hat{w}_1, ..., \hat{w}_L), \tag{16}$$

where $\hat{w}_j$ is the embedding of given instruction. Finally, we use two cross attention modules to implement landmark- and action-speaker:

$$\begin{aligned} \tilde{\gamma}_{t,j} &= \text{Sigmoid}\left((W_l \hat{h}_t^v)^T \hat{h}_j^l\right), \\ \tilde{\sigma}_{t,j} &= \text{Sigmoid}\left((W_a \hat{h}_t^v)^T \hat{h}_j^l\right), \end{aligned} \tag{17}$$

where $\tilde{\gamma}_{t,j}$ and $\tilde{\sigma}_{t,j}$ is the probability of j-th word that belongs to landmark and action related part which should be highlighted to navigate to the next viewpoint.

To train the DLS, we use the annotated label to optimize the Binary Cross Entropy loss. The loss formulation can be obtained by replacing $\gamma_{t,j}$ in Eq. (12) and $\sigma_{t,j}$ in Eq. (13) with $\tilde{\gamma}_{t,j}$ and $\tilde{\sigma}_{t,j}$, respectively. To train our full model, we firstly train a converged DLS, and then freeze its parameters whose output will be regarded as pseudo-labels to optimize the language attention weight.

3.6 Training

The model is trained by mixed Imitation Learning (IL) and Reinforcement Learning (RL). In IL phase, the agent can learn quickly from the teacher action a_t^* at each time step t by Behaviour Cloning [5]. The IL loss is formulated by: $\mathcal{L}_{IL} = \frac{1}{T}\sum_{t=1}^{T} -a_t^* \log(p_t)$. In RL phase, the agent learns from the rewards by taking the action a_t^s sampled with the probability p_t. Formally: $\mathcal{L}_{RL} = \frac{1}{T}\sum_{t=1}^{T} -a_t^s \log(p_t) A_t$. where A_t is the advantage in A2C algorithm [49]. Overall, we jointly train our model in an end-to-end manner using the loss formulation:

$$\mathcal{L}_{loss} = \mathcal{L}_{RL} + \lambda_1 \mathcal{L}_{IL} + \lambda_2 \mathcal{L}_{lan} + \lambda_3 \mathcal{L}_{act}. \quad (18)$$

where λ_1 manages the trade-off between IL and RL, λ_2 and λ_3 are weighting coefficients of language auxiliary loss.

4 Experiments

4.1 Experimental Setup

Evaluation Metrics. Following previous works [46,79], we use standard metrics to evaluate the navigator's performance on R2R dataset [2]. These include Success Rate (SR) which is a ratio of the agent whose distance between stopped position and target location is within 3 meters, Success rate weighted by Path Length (SPL), Navigation Error (NE) which is the average distance in meters between the final position and the target, and Oracle Success Rate (OSR) which measures success rate at the nearest point to the goal along the entire visited path. Among these metrics, SR and SPL are the main metrics, since the SR directly quantifies the crucial notion of success rate for the VLN task, and SPL combines the path length and SR to focus on more efficient navigation. For R4R [32], additional metrics including Coverage weighted by Length Score (CLS) [32], normalized Dynamic Time Warping (nDTW) and SDTW [30] are considered to encourage the agent to stay on the path that the instruction indicates.
Implementation Details. For the decoupled label speaker, the model is trained on the proposed LAR2R dataset for 80,000 iterations with a batch size of 32. The Adam [34] optimizer is used with a learning rate of 1e-4. Then the model with the lowest loss on the validation unseen set is selected. For navigation, we set the language auxiliary loss weights to $\lambda_2 = 1.0$, $\lambda_3 = 1.0$. We keep the other settings same as the baseline [56,28,11] for fairness.

4.2 Results and Analysis

Comparison to SoTA. The single-run setting is considered as the primary experimental setup since it can accurately reflect the agent's performance and generalizability to novel environments and instructions. Under this setting, the agent is not allowed

Model	R2R Validation unseen				R2R Test unseen			
	SR↑	**SPL↑**	OR↑	NE↓	**SR↑**	**SPL↑**	OR↑	NE↓
Speaker-Follower [18]	36	-	45	6.62	35	28	44	6.62
RCM [73]	43	-	50	6.09	43	38	50	6.12
Self-Monitoring [46]	45	32	56	5.52	43	32	55	5.99
Regretful [47]	50	41	59	5.32	48	40	56	5.69
E-Dropout [65]	52	48	-	5.22	51	47	59	5.23
AuxRN [79]	55	50	62	5.28	55	51	62	5.15
OAAM [56]	54	50	61	-	53	50	61	-
RelGraph [26]	57	53	-	4.73	55	52	-	4.75
PREVALENT [23]	58	53	-	4.71	54	51	-	5.30
SSM [70]	62	45	73	4.32	61	46	70	4.57
VLN↻BERT [28]	63	57	-	3.93	63	57	-	4.09
HAMT [11]	66	61	-	**2.29**	65	60	-	3.93
OAAM* [56]	54.4	49.0	62.8	5.00	53.6	49.9	59.4	5.00
OAAM* + DDL	57.6	51.0	65.6	4.63	57.0	51.4	65.3	4.70
VLN↻BERT* [28]	62.2	56.5	68.3	4.09	62.2	56.7	68.6	4.04
VLN↻BERT* + DDL	64.8	58.3	71.1	3.84	64.1	58.1	70.8	3.97
HAMT* [11]	65.6	60.7	73.7	3.51	64.4	59.5	69.3	4.03
HAMT* + DDL	**67.9**	**62.2**	**76.0**	3.38	**66.3**	**61.1**	**72.4**	**3.80**

Table 1. Comparison of single-run performance to the state-of-the-art methods on R2R [2]. *denotes our re-implementation. DDL provides consistent improvements.

to run multiple trials or pre-explore the test environments. As shown in Table 1, DDL brings consistent and substantial performance improvement to both the LSTM-based and BERT-based navigators, demonstrating the generality and effectiveness of our approach. For the state-of-the-art method HAMT, DDL increases the success rate by 2.3% and SPL with 1.5% on validation unseen set. On test unseen, we increase SR by 1.9%, while SPL is improved by 1.6%. Table 2 shows we can also boost the performance on R4R in terms of nDTW, SDTW and CLS, indicating that DDL can encourage the agent to stay on the path and have high instruction fidelity.

Ablation Study. Table 3 presents the impact of each component in OAAM. The training process consists of two stages. In the first stage, only the original training data is used. Thus, we use the annotated labels for IL phase and pseudo-labels for RL phase. In the second stage, a large amount of augmented data is added, which is unlabeled. We utilize Decoupled Label Speaker to provide intermediate supervision signals for this augmented data. As shown in Table 3, we find that when the annotated labels (model #2) are used to regularize the language attention weight in IL phase, the performance gets slight improvement. Moreover, model #3 indicates that the generalizability of the agent can be improved via providing landmark and action pseudo-labels for the reinforcement training phase. Comparing model #5 with #4, we find that the performance on validation unseen split gets significant improvement with the gains of 3.2% and 2.0% in terms of SR and SPL. This can be attributed to the fact that although Back Translation (BT) [65] brings lots of data without decoupled labels, our speaker can accurately generate landmark and action pseudo-labels for the augmented data thereby providing additional supervision signals during training.

Effectiveness of Decoupled Label. Based on the proposed LAR2R labels, initially, we only utilize the annotated label to regularize language attention weight in IL phase. Under this setting, to reveal its effectiveness, Table 4 shows the results of different types

Model	R4R Validation seen						R4R Validation unseen					
	SR↑	SPL↑	NE↓	**nDTW↑**	**SDTW↑**	**CLS↑**	SR↑	SPL↑	NE↓	**nDTW↑**	**SDTW↑**	**CLS↑**
Speaker-Follower [18]	52	37	5.35	-	-	46	24	12	8.47	-	-	30
RCM [32]	53	31	5.37	-	-	55	26	8	8.08	-	-	35
PTA [39]	58	39	4.53	**58**	41	**60**	24	10	8.25	32	10	37
EnvDrop [65]	52	41	-	-	27	53	29	18	-	-	9	34
EGP [13]	-	-	-	-	-	-	30	-	8.00	37	18	**44**
OAAM* [56]	48.3	40.2	5.81	47.6	31.2	51.0	26.6	19.0	8.51	30.3	12.6	36.2
OAAM* + DDL	50.2	41.9	5.59	49.8	33.6	53.7	28.5	21.2	8.15	33.1	14.2	38.5
VLN↻BERT* [28]	60.2	50.7	4.63	48.2	36.3	49.5	39.3	29.3	6.66	35.2	19.1	39.4
VLN↻BERT* + DDL	**64.4**	**53.6**	**3.97**	55.6	**43.1**	57.6	**42.4**	**32.7**	**6.43**	**38.5**	**21.0**	**43.6**

Table 2. Comparison of single-run performance to the state-of-the-art methods on R4R [32]. *denotes our re-implementation. DDL provides consistent improvements.

Model	Component				R2R Val seen		R2R Val unseen	
	baseline	LAR2R	DLS	BT	SR↑	SPL↑	SR↑	SPL↑
1	✓				63.0	59.5	50.2	45.4
2	✓	✓			65.3	61.1	50.8	45.7
3	✓	✓	✓		65.2	61.4	51.5	45.9
4	✓			✓	70.7	67.1	54.4	49.0
5	✓	✓	✓	✓	70.8	66.4	**57.6**	**51.0**

Table 3. Ablation study with OAAM showing the effect of each component on R2R. LAR2R means the annotated labels, DLS represents the pseudo-labels, and BT denotes extra augmented training data [65] without decoupled labels.

of language label. In the first column, *Random* represents the language labels $x_{t,j}$ and $y_{t,j}$ in Eqs. (12) and (13) are sampled from a uniform distribution $U[0,1]$. *Average* means all language labels are assigned to 1. *FGR2R* means we generate the labels by Part-of-Speech tagging for each sub-instruction of FGR2R [27]. As shown in Table 4, the random label degrades the performance with the reduction of 2.3% SR and 1.2% SPL on validation unseen set compared with model #1 in Table 3 without language label. Moreover, our proposed labels have better performance than that of FGR2R, since FGR2R only focuses on the segmentation while ours is more fine-grained. These results further demonstrate the effectiveness of our decoupled labels.

Quantitative and Qualitative Analysis of Decoupled Label Speaker. Fig. 5 presents an example of the distribution of landmark and action attention weights predicted by our DLS at two navigation steps. One can note that the landmark- and action-related instructions are clearly disentangled. In particular, the landmark-speaker not only focuses on the object (*e.g.* bed), but it is also able to attend the specific position next to the object (*e.g.* the end of the bed), which can help the agent navigate to the precise location. Notice that the example is tested in an unseen environment, demonstrating the generalizability of our model. Moreover, a quantitative analysis is presented in Supp-Figure 1. It can be noted that most pseudo-labels have high cosine similarity with the human-annotated labels, showing the effectiveness of DLS.

Model	R2R Val seen				R2R Val unseen			
	SR↑	**SPL↑**	OR↑	NE↓	**SR↑**	**SPL↑**	OR↑	NE↓
Random	53.6	50.0	60.9	5.03	47.9	44.2	55.6	5.45
Average	62.3	58.3	68.0	4.19	48.7	45.3	55.8	5.43
*FGR2R** [27]	63.0	58.8	71.2	4.09	49.6	44.6	56.9	5.41
Ours (LAR2R)	**65.3**	**61.1**	**72.4**	**3.84**	**50.8**	**45.7**	**58.1**	**5.27**

Table 4. Performance comparison with OAAM considering different types of decoupled labels on R2R. Our fine-grained labels perform favorably against the other alternatives.

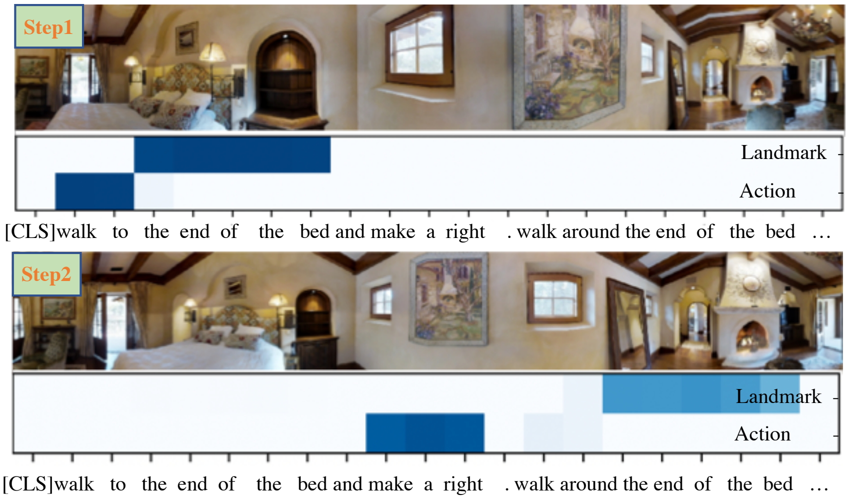


Fig. 5. Distribution of landmark and action attention weights predicted by the decoupled label speaker at the first two navigation steps in an unseen environment. Color shade represents the relative attention weight (darker is higher).

5 Conclusion

In this paper, we have explored the effectiveness of the decoupled instruction label on the vision-and-language navigation task. Firstly, we enrich R2R with specific landmark- and action-aware labels. We further propose a Decoupled Label Speaker to generate pseudo-labels, which are utilized to guide discriminative feature extraction in Disentangled Decoding Module. Superior performance on two VLN benchmarks demonstrates the effectiveness of our proposed approach. Although this work focuses on using the decoupled labels to provide accurate inputs for VLN, this framework can positively impact other tasks, such as visual question answering [3,77,63] and visual dialog navigation [51,52,67]. Further, new solutions for achieving disentanglement is an critical open research question in VLN, as well as other computer vision tasks, such as object tracking [15,22] and segmentation [75,16].

Acknowledgements. This work was supported partly by the Start-up Research Grant (SRG) of University of Macau. We thank the reviewers for their insightful comments.

References

1. Anderson, P., He, X., Buehler, C., Teney, D., Johnson, M., Gould, S., Zhang, L.: Bottom-up and top-down attention for image captioning and visual question answering. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 6077–6086 (2018) 5
2. Anderson, P., Wu, Q., Teney, D., Bruce, J., Johnson, M., Sünderhauf, N., Reid, I., Gould, S., Van Den Hengel, A.: Vision-and-language navigation: Interpreting visually-grounded navigation instructions in real environments. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 3674–3683 (2018) 2, 3, 11, 12
3. Antol, S., Agrawal, A., Lu, J., Mitchell, M., Batra, D., Zitnick, C.L., Parikh, D.: Vqa: Visual question answering. In: Proceedings of the IEEE International Conference on Computer Vision (December 2015) 14
4. Ba, J.L., Kiros, J.R., Hinton, G.E.: Layer normalization. arXiv preprint arXiv:1607.06450 (2016) 8
5. Bojarski, M., Del Testa, D., Dworakowski, D., Firner, B., Flepp, B., Goyal, P., Jackel, L.D., Monfort, M., Muller, U., Zhang, J., et al.: End to end learning for self-driving cars. arXiv preprint arXiv:1604.07316 (2016) 4, 11
6. Cao, K., Brbić, M., Leskovec, J.: Concept learners for few-shot learning. In: International Conference on Learning Representations (2021) 2
7. Chang, A., Dai, A., Funkhouser, T., Halber, M., Niebner, M., Savva, M., Song, S., Zeng, A., Zhang, Y.: Matterport3d: Learning from rgb-d data in indoor environments. In: 7th IEEE International Conference on 3D Vision, 3DV 2017. pp. 667–676. Institute of Electrical and Electronics Engineers Inc. (2018) 3
8. Chen, H., Suhr, A., Misra, D., Snavely, N., Artzi, Y.: Touchdown: Natural language navigation and spatial reasoning in visual street environments. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12538–12547 (2019) 4
9. Chen, J., Gao, C., Meng, E., Zhang, Q., Liu, S.: Reinforced structured state-evolution for vision-language navigation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 15450–15459 (June 2022) 4
10. Chen, K., Chen, J.K., Chuang, J., Vázquez, M., Savarese, S.: Topological planning with transformers for vision-and-language navigation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11276–11286 (2021) 4
11. Chen, S., Guhur, P.L., Schmid, C., Laptev, I.: History aware multimodal transformer for vision-and-language navigation. Advances in Neural Information Processing Systems **34** (2021) 3, 4, 5, 6, 11, 12
12. Chen, S., Guhur, P.L., Tapaswi, M., Schmid, C., Laptev, I.: Think global, act local: Dual-scale graph transformer for vision-and-language navigation. arXiv preprint arXiv:2202.11742 (2022) 4
13. Deng, Z., Narasimhan, K., Russakovsky, O.: Evolving graphical planner: Contextual global planning for vision-and-language navigation. In: Advances in Neural Information Processing Systems. vol. 33, pp. 20660–20672. Curran Associates, Inc. (2020) 1, 13
14. Devlin, J., Chang, M., Lee, K., Toutanova, K.: BERT: pre-training of deep bidirectional transformers for language understanding. 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). pp. 4171–4186 (2019) 4

15. Dong, X., Shen, J., Shao, L., Porikli, F.: Clnet: A compact latent network for fast adjusting siamese trackers. In: European Conference on Computer Vision. pp. 378–395. Springer (2020) 14
16. Dong, X., Shen, J., Shao, L., Van Gool, L.: Sub-markov random walk for image segmentation. IEEE Transactions on Image Processing **25**(2), 516–527 (2015) 14
17. Fang, K., Toshev, A., Fei-Fei, L., Savarese, S.: Scene memory transformer for embodied agents in long-horizon tasks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 538–547 (2019) 4
18. Fried, D., Hu, R., Cirik, V., Rohrbach, A., Andreas, J., Morency, L.P., Berg-Kirkpatrick, T., Saenko, K., Klein, D., Darrell, T.: Speaker-follower models for vision-and-language navigation. In: Proceedings of the 32nd International Conference on Neural Information Processing Systems. pp. 3318–3329 (2018) 1, 3, 4, 12, 13
19. Fu, T.J., Wang, X.E., Peterson, M.F., Grafton, S.T., Eckstein, M.P., Wang, W.Y.: Counterfactual vision-and-language navigation via adversarial path sampler. In: European Conference on Computer Vision. pp. 71–86. Springer (2020) 1
20. Gao, C., Chen, J., Liu, S., Wang, L., Zhang, Q., Wu, Q.: Room-and-object aware knowledge reasoning for remote embodied referring expression. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 3064–3073 (June 2021) 4
21. Guhur, P.L., Tapaswi, M., Chen, S., Laptev, I., Schmid, C.: Airbert: In-domain pretraining for vision-and-language navigation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 1634–1643 (2021) 1
22. Han, W., Dong, X., Khan, F.S., Shao, L., Shen, J.: Learning to fuse asymmetric feature maps in siamese trackers. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 16570–16580 (2021) 14
23. Hao, W., Li, C., Li, X., Carin, L., Gao, J.: 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. pp. 13137–13146 (2020) 1, 12
24. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016) 5
25. Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural computation **9**(8), 1735–1780 (1997) 10
26. Hong, Y., Rodriguez, C., Qi, Y., Wu, Q., Gould, S.: Language and visual entity relationship graph for agent navigation. Advances in Neural Information Processing Systems **33** (2020) 2, 4, 12
27. Hong, Y., Rodriguez, C., Wu, Q., Gould, S.: Sub-instruction aware vision-and-language navigation. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing. pp. 3360–3376 (2020) 2, 13, 14
28. Hong, Y., Wu, Q., Qi, Y., Rodriguez-Opazo, C., Gould, S.: Vln bert: A recurrent vision-and-language bert for navigation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1643–1653 (June 2021) 3, 4, 5, 11, 12, 13
29. Hu, R., Fried, D., Rohrbach, A., Klein, D., Darrell, T., Saenko, K.: Are you looking? grounding to multiple modalities in vision-and-language navigation. arXiv preprint arXiv:1906.00347 (2019) 4

30. Ilharco, G., Jain, V., Ku, A., Ie, E., Baldridge, J.: General evaluation for instruction conditioned navigation using dynamic time warping. arXiv preprint arXiv:1907.05446 (2019) 11
31. Irshad, M.Z., Mithun, N.C., Seymour, Z., Chiu, H.P., Samarasekera, S., Kumar, R.: Sasra: Semantically-aware spatio-temporal reasoning agent for vision-and-language navigation in continuous environments. arXiv preprint arXiv:2108.11945 (2021) 4
32. Jain, V., Magalhaes, G., Ku, A., Vaswani, A., Ie, E., Baldridge, J.: Stay on the path: Instruction fidelity in vision-and-language navigation. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. pp. 1862–1872 (2019) 1, 3, 11, 13
33. Ke, L., Li, X., Bisk, Y., Holtzman, A., Gan, Z., Liu, J., Gao, J., Choi, Y., Srinivasa, S.: Tactical rewind: Self-correction via backtracking in vision-and-language navigation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6741–6749 (2019) 1, 3
34. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. In: 3rd International Conference on Learning Representations, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings (2015) 11
35. Kolve, E., Mottaghi, R., Han, W., VanderBilt, E., Weihs, L., Herrasti, A., Gordon, D., Zhu, Y., Gupta, A., Farhadi, A.: Ai2-thor: An interactive 3d environment for visual ai. arXiv preprint arXiv:1712.05474 (2017) 3
36. Krantz, J., Gokaslan, A., Batra, D., Lee, S., Maksymets, O.: Waypoint models for instruction-guided navigation in continuous environments. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 15162–15171 (2021) 4
37. Krantz, J., Wijmans, E., Majumdar, A., Batra, D., Lee, S.: Beyond the nav-graph: Vision-and-language navigation in continuous environments. In: European Conference on Computer Vision. pp. 104–120. Springer (2020) 4
38. Ku, A., Anderson, P., Patel, R., Ie, E., Baldridge, J.: 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. pp. 4392–4412 (2020) 2, 4
39. Landi, F., Baraldi, L., Cornia, M., Corsini, M., Cucchiara, R.: Perceive, transform, and act: Multimodal attention networks for low-level vision-and-language navigation. arXiv preprint arXiv:1911.12377 (2019) 4, 13
40. Li, X., Li, C., Xia, Q., Bisk, Y., Celikyilmaz, A., Gao, J., Smith, N.A., Choi, Y.: Robust navigation with language pretraining and stochastic sampling. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing. pp. 1494–1499 (2019) 1, 4
41. Li, X., Yin, X., Li, C., Zhang, P., Hu, X., Zhang, L., Wang, L., Hu, H., Dong, L., Wei, F., et al.: Oscar: Object-semantics aligned pre-training for vision-language tasks. In: European Conference on Computer Vision. pp. 121–137. Springer (2020) 1
42. Liang, X., Zhu, F., Zhu, Y., Lin, B., Wang, B., Liang, X.: Contrastive instruction-trajectory learning for vision-language navigation. arXiv preprint arXiv:2112.04138 (2021) 4
43. Lin, C., Jiang, Y., Cai, J., Qu, L., Haffari, G., Yuan, Z.: Multimodal transformer with variable-length memory for vision-and-language navigation. arXiv preprint arXiv:2111.05759 (2021) 4

44. Liu, C., Zhu, F., Chang, X., Liang, X., Ge, Z., Shen, Y.D.: Vision-language navigation with random environmental mixup. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 1644–1654 (2021) 3, 9
45. Lu, J., Batra, D., Parikh, D., Lee, S.: Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In: Advances in Neural Information Processing Systems. vol. 32. Curran Associates, Inc. (2019) 1
46. Ma, C.Y., Lu, J., Wu, Z., AlRegib, G., Kira, Z., Socher, R., Xiong, C.: Self-monitoring navigation agent via auxiliary progress estimation. In: Proceedings of the International Conference on Learning Representations (2019) 1, 3, 4, 11, 12
47. Ma, C.Y., Wu, Z., AlRegib, G., Xiong, C., Kira, Z.: The regretful agent: Heuristic-aided navigation through progress estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6732–6740 (2019) 1, 3, 12
48. Majumdar, A., Shrivastava, A., Lee, S., Anderson, P., Parikh, D., Batra, D.: Improving vision-and-language navigation with image-text pairs from the web. In: European Conference on Computer Vision. pp. 259–274. Springer (2020) 1, 4
49. Mnih, V., Badia, A.P., Mirza, M., Graves, A., Lillicrap, T., Harley, T., Silver, D., Kavukcuoglu, K.: Asynchronous methods for deep reinforcement learning. In: International conference on machine learning. pp. 1928–1937. PMLR (2016) 11
50. Moudgil, A., Majumdar, A., Agrawal, H., Lee, S., Batra, D.: Soat: A scene-and object-aware transformer for vision-and-language navigation. arXiv preprint arXiv:2110.14143 (2021) 4
51. Nguyen, K., Daumé III, H.: Help, anna! visual navigation with natural multimodal assistance via retrospective curiosity-encouraging imitation learning. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing. pp. 684–695 (2019) 3, 14
52. Nguyen, K., Dey, D., Brockett, C., Dolan, B.: Vision-based navigation with language-based assistance via imitation learning with indirect intervention. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12527–12537 (2019) 3, 14
53. Parvaneh, A., Abbasnejad, E., Teney, D., Shi, Q., van den Hengel, A.: Counterfactual vision-and-language navigation: Unravelling the unseen. Advances in Neural Information Processing Systems **33** (2020) 1
54. Pashevich, A., Schmid, C., Sun, C.: Episodic Transformer for Vision-and-Language Navigation. In: ICCV (2021) 4
55. Qi, Y., Pan, Z., Hong, Y., Yang, M.H., van den Hengel, A., Wu, Q.: The road to know-where: An object-and-room informed sequential bert for indoor vision-language navigation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 1655–1664 (2021) 4
56. Qi, Y., Pan, Z., Zhang, S., van den Hengel, A., Wu, Q.: Object-and-action aware model for visual language navigation. In: Proceedings of the European Conference on Computer Vision, Glasgow, Scotland. pp. 23–28. Springer (2020) 1, 2, 3, 4, 5, 7, 11, 12, 13
57. Qi, Y., Wu, Q., Anderson, P., Wang, X., Wang, W.Y., Shen, C., Hengel, A.v.d.: Reverie: Remote embodied visual referring expression in real indoor environments. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9982–9991 (2020) 3
58. Qiao, Y., Qi, Y., Hong, Y., Yu, Z., Wang, P., Wu, Q.: Hop: History-and-order aware pre-training for vision-and-language navigation. In: Proceedings of the IEEE/CVF

Conference on Computer Vision and Pattern Recognition (CVPR). pp. 15418–15427 (June 2022) 4

59. Qin, W., Misu, T., Wijaya, D.: Explore the potential performance of vision-and-language navigation model: a snapshot ensemble method. arXiv preprint arXiv:2111.14267 (2021) 4
60. Raychaudhuri, S., Wani, S., Patel, S., Jain, U., Chang, A.X.: Language-aligned waypoint (law) supervision for vision-and-language navigation in continuous environments. arXiv preprint arXiv:2109.15207 (2021) 4
61. Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., et al.: Imagenet large scale visual recognition challenge. International Journal of Computer Vision **115**(3) (2015) 5
62. Savva, M., Kadian, A., Maksymets, O., Zhao, Y., Wijmans, E., Jain, B., Straub, J., Liu, J., Koltun, V., Malik, J., et al.: Habitat: A platform for embodied ai research. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 9339–9347 (2019) 3
63. Shih, K.J., Singh, S., Hoiem, D.: Where to look: Focus regions for visual question answering. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (June 2016) 14
64. Shridhar, M., Thomason, J., Gordon, D., Bisk, Y., Han, W., Mottaghi, R., Zettlemoyer, L., Fox, D.: Alfred: A benchmark for interpreting grounded instructions for everyday tasks. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10740–10749 (2020) 3
65. Tan, H., Yu, L., Bansal, M.: 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). pp. 2610–2621 (2019) 1, 3, 4, 5, 9, 12, 13
66. Tan, S., Ge, M., Guo, D., Liu, H., Sun, F.: Self-supervised 3d semantic representation learning for vision-and-language navigation. arXiv preprint arXiv:2201.10788 (2022) 4
67. Thomason, J., Murray, M., Cakmak, M., Zettlemoyer, L.: Vision-and-dialog navigation. In: Conference on Robot Learning. pp. 394–406. PMLR (2020) 3, 14
68. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I.: Attention is all you need. In: Proceedings of the 31st International Conference on Neural Information Processing Systems. pp. 6000–6010 (2017) 2, 4
69. Wang, H., Liang, W., Shen, J., Van Gool, L., Wang, W.: Counterfactual cycle-consistent learning for instruction following and generation in vision-language navigation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 15471–15481 (2022) 1, 4
70. Wang, H., Wang, W., Liang, W., Xiong, C., Shen, J.: Structured scene memory for vision-language navigation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8455–8464 (2021) 4, 12
71. Wang, H., Wang, W., Shu, T., Liang, W., Shen, J.: Active visual information gathering for vision-language navigation. In: European Conference on Computer Vision. pp. 307–322. Springer (2020) 1, 4
72. Wang, H., Wu, Q., Shen, C.: Soft expert reward learning for vision-and-language navigation. In: European Conference on Computer Vision (2020) 1, 4
73. Wang, X., Huang, Q., Celikyilmaz, A., Gao, J., Shen, D., Wang, Y.F., Wang, W.Y., Zhang, L.: Reinforced cross-modal matching and self-supervised imitation learning

for vision-language navigation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6629–6638 (2019) 1, 4, 12

74. Wang, X., Xiong, W., Wang, H., Wang, W.Y.: Look before you leap: Bridging model-free and model-based reinforcement learning for planned-ahead vision-and-language navigation. In: Proceedings of the European Conference on Computer Vision. pp. 37–53 (2018) 1, 4
75. Wu, D., Dong, X., Shao, L., Shen, J.: Multi-level representation learning with semantic alignment for referring video object segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4996–5005 (2022) 14
76. Xiang, J., Wang, X., Wang, W.Y.: Learning to stop: A simple yet effective approach to urban vision-language navigation. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings. pp. 699–707 (2020) 1
77. Zhang, Y., Niebles, J.C., Soto, A.: Interpretable visual question answering by visual grounding from attention supervision mining. In: 2019 ieee winter conference on applications of computer vision. pp. 349–357. IEEE (2019) 14
78. Zhou, B., Lapedriza, A., Khosla, A., Oliva, A., Torralba, A.: Places: A 10 million image database for scene recognition. IEEE transactions on pattern analysis and machine intelligence **40**(6), 1452–1464 (2017) 5
79. Zhu, F., Zhu, Y., Chang, X., Liang, X.: Vision-language navigation with self-supervised auxiliary reasoning tasks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10012–10022 (2020) 1, 4, 11, 12
80. Zhu, W., Hu, H., Chen, J., Deng, Z., Jain, V., Ie, E., Sha, F.: Babywalk: Going farther in vision-and-language navigation by taking baby steps. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. pp. 2539–2556 (2020) 2