Vision-and-Language Navigation via Causal Learning

Liuyi Wang, Zongtao He, Ronghao Dang, Mengjiao Shen, Chengju Liu*, Qijun Chen*
School of Electronic and Information Engineering, Tongji University, Shanghai, China
{wly, xingchen327, dangronghao, liuchengju, qjchen}@tongji.edu.cn, mojo.shum@outlook.com

Abstract

*In the pursuit of robust and generalizable environment perception and language understanding, the ubiquitous challenge of dataset bias continues to plague vision-and-language navigation (VLN) agents, hindering their performance in unseen environments. This paper introduces the **g**eneralized cr**o**ss-modal c**a**usal **t**ransformer (GOAT), a pioneering solution rooted in the paradigm of causal inference. By delving into both observable and unobservable confounders within vision, language, and history, we propose the **b**ack-door and **f**ront-door **a**djustment **c**ausal **l**earning (BACL and FACL) modules to promote unbiased learning by comprehensively mitigating potential spurious correlations. Additionally, to capture global confounder features, we propose a **c**ross-modal **f**eature **p**ooling (CFP) module supervised by contrastive learning, which is also shown to be effective in improving cross-modal representations during pre-training. Extensive experiments across multiple VLN datasets (R2R, REVERIE, RxR, and SOON) underscore the superiority of our proposed method over previous state-of-the-art approaches. Code is available at https://github.com/CrystalSixone/VLN-GOAT.*

1. Introduction

Effective environment perception, language understanding, and historical utilization are at the core of vision-and-language navigation (VLN) [4]. Despite significant progress, deploying VLN in the real world remains a huge challenge, primarily due to diversities and uncertainties in environments and instructions. A key hindrance is dataset bias [79, 82], *e.g.*, agents may overfit to familiar visual environments, resulting in diminished performance in environments with diverse appearances and layouts [24]. This over-reliance on specific patterns, like biased structural trajectories and repeated entity components, raises concerns about the robustness and generalizability of VLN systems.

One way to mitigate dataset bias in VLN is to build broader and more diverse datasets, which is what numerous recent studies have focused on. These include using speaker models to generate pseudo instructions [14, 16, 61, 62, 64, 65], synthesizing cross-environmental trajectories [26, 39], transferring image styles [30, 34], collecting data from the web [18, 36, 44, 70], and labeling more fine-grained, entity-aligned instructions [10, 21, 29]. However, achieving a perfectly balanced dataset devoid of bias is nearly impossible. Consequently, we often find ourselves caught in a cycle of "creating a dataset" - "identifying bias" - "creating a new dataset". Therefore, we are prompted to move from diagnosis to treatment [73], transitioning from the continuous collection of new datasets to the development of unbiased models that can confront and mitigate bias.

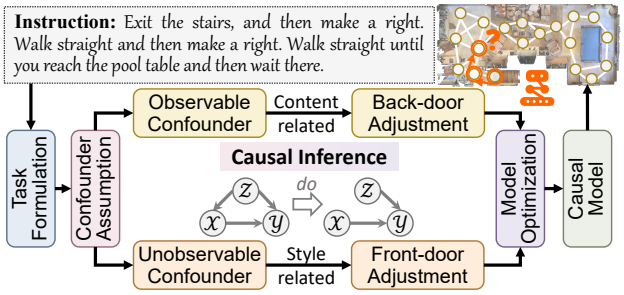

Figure 1. In response to language instructions, the VLN agent is required to navigate to the target based on visual cues. This paper introduces a causal learning pipeline using the *do*-operator to reduce bias from confounders in VLN action prediction.

However, existing methods that focus on model designing mainly concentrate on introducing more types of inputs (*e.g.*, objects [2, 12, 13, 23, 35, 63] and depth [3, 20, 40, 77]), or constructing the global graph [8, 9, 46] to represent environments. These efforts, although valuable, often overlook underlying dataset biases and the essential causal logic behind the task. In fact, the reason why humans can well execute various instructions and navigate in unknown environments is that we can learn the inherent causality of events beyond biased observation, achieving good analogical association capability. Therefore, for the first time, we propose to use causal inference [51] to equip VLN agents with similar cognitive abilities that we have, and then allow

*Corresponding author.

them to make more reasoned decisions.

Then, how to develop such a causal inference capability? Although there is no single answer, we propose to exploit the concepts of *intervention* [52] – technique that uses a *do*-operator to alleviate the negative effects raised by confounders. Here, confounders are variables that influence both inputs and outcomes, creating spurious correlations and biases. Intervention empowers researchers to mitigate the impact of confounders, enabling the model to grasp the causation of events during data fitting. However, given the fact that VLN is such a complex task that involves cross-modal inputs and a long-term decision-making process, it is challenging to identify underlying confounders and apply intervention to debias through network learning.

To address the above challenges, we propose a **g**eneralized cr**o**ss-modal c**a**usal **t**ransformer (GOAT) approach that enables the VLN model to alleviate the negative effects raised by confounders, thus achieving causal decision-making (Fig. 1). Firstly, we propose a unified structural causal model to describe the VLN system, involving two distinct categories of confounders: *observable* and *unobservable*. Observable confounders are content-related and easily identifiable (*e.g.*, the keywords in instructions and the room references in environments), whereas unobservable confounders refer to intricate stylistic nuances that are harder to discern but can impact the overall system (*e.g.*, decoration styles in vision, sentence patterns in language, and trajectory trends in history). Then, we propose to address these confounders via two causal learning modules that are based on back-door and front-door adjustments [52] (namely BACL and FACL), respectively. Furthermore, to build global dictionaries for representing confounders, we devise a cross-modal feature pooling (CFP) module to effectively aggregate long-sequential features. Contrastive learning [56] is adopted to optimize CFP, serving as an additional auxiliary task during pre-training. As demonstrated by thorough experiments, our findings reveal the impact of integrating causal learning to deconfound biases on cross-modal inputs, offering valuable insights for enhancing generalization in similar tasks across diverse scenarios.

To summarize, our main contributions are as follows:

- We propose a unified structural causal model for VLN by comprehensively considering the observable and unobservable confounders hidden in different modalities.
- we propose BACL and FACL, using the back-door and front-door adjustments to allow end-to-end unbiased cross-modal intervention and decision-making.
- We propose CFP, a cross-modal feature pooling module designed to aggregate sequence features for semantic alignment and confounder dictionaries construction.
- Our GOAT model demonstrates exceptional generalization across multiple VLN datasets (R2R [4], RxR [29], REVERIE [53], and SOON [81]), outperforming existing state-of-the-art methods. A comprehensive causal learning pipeline is presented to inspire future research.

2. Related Work

Vision-and-Language Navigation (VLN) requires agents to navigate to specific locations [4, 29] or find target objects [53, 81] in real visual environments based on natural instructions. Its practicability has led to significant interest, showing potential in fundamental embodied AI skills. Initial models relied on recurrent neural networks [2, 11, 21, 68]. Transformer-based models [8, 9, 19, 22, 63] brought substantial progress due to their powerful long-distance encoding. However, small-scale datasets in VLN were found to cause bias, leading to serious overfitting. Consequently, several approaches were devised to tackle this challenge. Speaker-follower frameworks [16, 27, 58, 62, 64] used the speaker model to generate pseudo instructions. VLN-BERT [44], AirBERT [18], and Lily [36] collected trajectory-instruction pairs from diverse sources. REM [39] and EnvEdit [34] created new environments by editing existing environments. Although these methods improve generalization, they cannot completely eliminate inherent dataset biases. Therefore, we propose to develop an unbiased model by equipping the agent with the causal inference capability to learn cause-effect relations during data fitting, enabling them to adapt adeptly to diverse situations.

Causal Inference is an emerging technique exploring task causality [52], leading to a surge in efforts to integrate it with deep learning, in tasks like image recognition [67, 69, 76], image captioning [38, 73], and visual question answering [33, 49]. One popular way is to use the adjustment technique to alleviate the negative effects caused by confounders, and some studies exploring the use of counterfactuals [1, 49, 50]. This paper emphasizes the adjustment method due to its practicality. However, most of the existing causal learning tasks are simple without considering more challenging tasks like VLN. Additionally, current methods apply back-door [38, 66, 75, 78] or front-door [42, 73, 74] adjustments separately across modalities, lacking comprehensive confounder assumptions and complete bias corrections. In this paper, we propose to simultaneously tackle both observable and unobservable confounders in vision, language, and history. This approach significantly reduces overall bias, enhancing the generalization capabilities of embodied VLN agents.

3. Preliminary

3.1. Task Formulation

The VLN task [4] involves an embodied agent following natural language instructions to navigate real indoor environments. Matterport3D simulator [7] is used to allow interaction, where the environment is provided as graphs

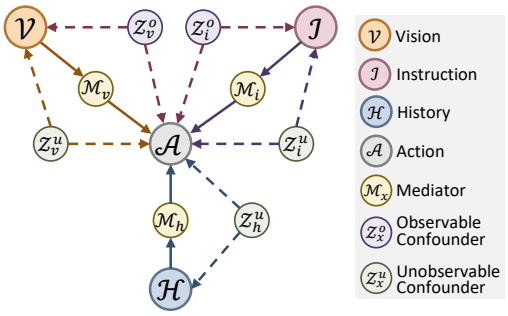

Figure 2. Illustration of the structural causal model of VLN.

with connected navigable nodes. The agent receives natural language instructions $\mathcal{I} = \{w_1, w_2, ..., w_L\}$ with L words, and the current panorama separated into 36 sub-images $\mathcal{V} = \{v_1, v_2, ..., v_{36}\}$ [16]. The agent also knows its current heading θ and elevation ϕ. During navigation, the agent needs to select the next point from nearby candidates or predicts the `stop` signal based on visual cues. Success is defined when the stop location is within 3 meters of the ground-truth position. For the goal-oriented task, REVERIE [53] and SOON [81] additionally require locating the target object at the final destination.

3.2. Structural Causal Model of VLN

As illustrated in Fig. 2, we construct a structural causal model capturing the relationships among the key variables in VLN: visual observation $\mathcal{V}$, linguistic instruction $\mathcal{I}$, decision history $\mathcal{H}$, and action prediction $\mathcal{A}$. To clarify, we use $\mathcal{X}$ to denote inputs ($\mathcal{X} = \{\mathcal{V}, \mathcal{I}, \mathcal{H}\}$) and $\mathcal{Y}$ for output ($\mathcal{Y} = \mathcal{A}$). In this directed acyclic graph, the starting and ending points represent the cause and effect, respectively. Traditional VLN methods focus on learning the observational association $P(\mathcal{Y}|\mathcal{X})$, overlooking the ambiguity introduced by confounders $\mathcal{Z}$ in the back-door path $\mathcal{X} \leftarrow \mathcal{Z} \rightarrow \mathcal{Y}$. Here, confounders are extraneous variables that influence both causes and effects, e.g., frequently occurring content or specific attributes. $\mathcal{Z} \rightarrow \mathcal{X}$ arises since the combined probability of input data is inevitably affected by the limited resources available in the real world when collection and simulation. Additionally, $\mathcal{Z} \rightarrow \mathcal{Y}$ exists because collected environments, labeled instructions, or sampled trajectories also affect the probability of action distributions. These confounder links enable spurious shortcuts during training but can be detrimental in new situations.

We propose to distinguish hidden confounders from different modalities into *observable* and *unobservable* categories, enhancing our prior knowledge integration and the rationality of assumption. Concretely, observable confounders encompass instances that can be recognized (*e.g.*, room references z_v^o and guiding keywords z_i^o). In contrast, unobservable confounders consist of intricate patterns and style-related elements that are challenging to qualitatively describe (*e.g.*, decoration style in vision z_v^u, sentence pattern in language z_i^u, and trajectory trend in history z_h^u). Since we cannot explicitly model unobservable confounders $\mathcal{Z}^u$, the additional mediators $\mathcal{M}$ are inserted between $\mathcal{X}$ and $\mathcal{Y}$ to establish front-door paths $\mathcal{X} \rightarrow \mathcal{M} \rightarrow \mathcal{Y}$. Detailed adjustment methods are introduced in subsequent sections.

4. Methodology

The overview of the GOAT model is shown in Fig. 3. The proposed back-door and front-door adjustment causal learning modules are detailed in Sec. 4.1 and Sec. 4.2, respectively. The cross-modal feature pooling method is subsequently introduced in Sec. 4.3. Finally, a practical causal learning pipeline is presented in Sec. 4.4.

4.1. Observable Causal Inference

Back-door Adjustment Causal Learning (BACL). Based on Bayes's theorem, the typical observational likelihood is as $P(\mathcal{Y}|\mathcal{X}) = \sum_z P(\mathcal{Y}|\mathcal{X}, z)P(z|\mathcal{X})$, where $P(z|\mathcal{X})$ could bring biased weights. *Do*-operator [52] provides scientifically sound methods for determining causal effects by severing the back-door link between $\mathcal{Z}$ and $\mathcal{X}$. According to the invariance and independence rules [17], we have:

$$P(\mathcal{Y}|do(\mathcal{X})) = \sum_z P(\mathcal{Y}|do(\mathcal{X}), z)P(z|do(\mathcal{X})) \quad (1)$$

$$= \sum_z P(\mathcal{Y}|\mathcal{X}, z)P(z) \quad (2)$$

In such case, the intervention is achieved by blocking the back-door path $\mathcal{Z} \rightarrow \mathcal{X}$, making $\mathcal{X}$ have a fair opportunity to incorporate causality-related factors for prediction. Previous methods [42, 73, 78] used the NWGM approximation [6] to directly pursue causal learning to the final outputs. However, these methods limit the intervention only to the network's final Softmax layer, overlooking possible biased features in shallow layers. Since the conditional probability is implicit in pattern recognition made by the trained neural network [48], we release the target effect of causal hypothesis to *learned features* rather than merely *outputs*. Obtaining unbiased features leads to unbiased predictions. Consequently, the specific network module is formulated as $f(\boldsymbol{x}, \boldsymbol{z})$, and the implementation of Eq. (2) becomes:

$$\mathcal{B}(\boldsymbol{x}, \boldsymbol{z}) = \mathbb{E}_z[f(\boldsymbol{x}, \boldsymbol{z})] \quad (3)$$

The linear function is used as $f(\boldsymbol{x}, \boldsymbol{z}) = f_x(\boldsymbol{x}) + f_z(\boldsymbol{z})$. Then Eq. (3) becomes $f_x(\boldsymbol{x}) + \mathbb{E}_z[f_z(\boldsymbol{z})]$. To obtain $\mathbb{E}_z[f_z(\boldsymbol{z})]$, there are two prevalent approaches: statistic-based [42, 66] and attention-based methods [38, 73]:

$$\textbf{Statistic}: \mathbb{E}_z[f_z(\boldsymbol{z})] = \sum_i \frac{|z_i|}{\sum_j |z_j|} f_z(\boldsymbol{z}_i) \quad (4)$$

$$\textbf{Attention}: \mathbb{E}_z[f_z(\boldsymbol{z})] = \sum_i \frac{\exp(\boldsymbol{h}\boldsymbol{z}_i^T)}{\sum_j \exp(\boldsymbol{h}\boldsymbol{z}_j^T)} f_z(\boldsymbol{z}_i) \quad (5)$$

where $|z_i|$ denotes the number of z belonging to the i-th

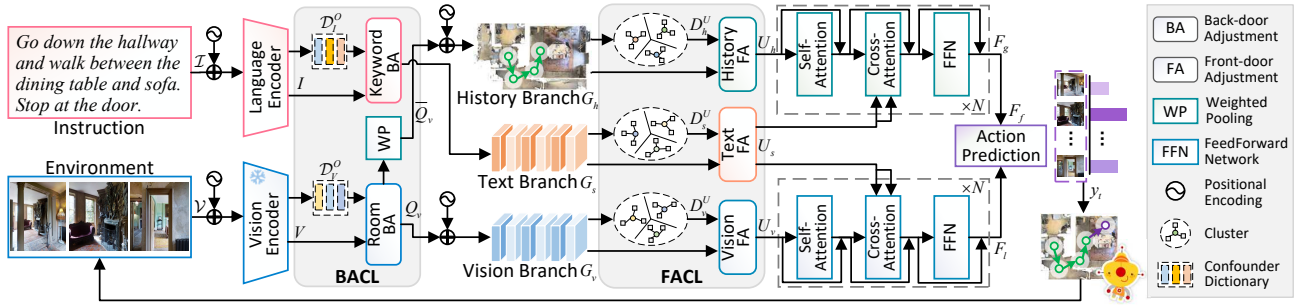

Figure 3. Framework of GOAT, built on the foundation of the dual-scale graph transformer [9]. Back-door and front-door adjustment causal learning mechanisms are used for mitigating spurious correlations, enabling unbiased feature learning and decision-making.

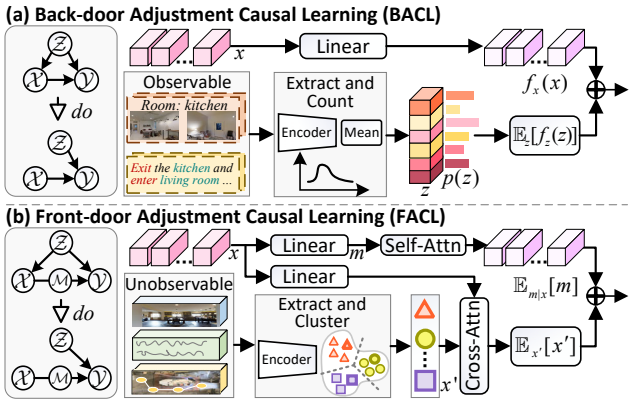

Figure 4. Illustration of BACL and FACL.

category in the training set, and $\boldsymbol{h}$ means hidden features. The illustration is shown in Fig. 4(a). Our experiments in Sec. 5.3 reveal modality-specific calculation preferences.

BACL in Text Content. In VLN instructions like "Exit the office and turn right into the kitchen," essential guiding elements such as directions (*e.g.*, "exit" and "right") and landmarks (*e.g.*, "office" and "kitchen") play significant roles. These keywords which are common causes of instruction construction and action distribution, serve as observable confounders. Firstly, we build the text keyword dictionary $\mathcal{D}_I^O = [\boldsymbol{z}_{i,1}^o, \boldsymbol{z}_{i,2}^o, ..., \boldsymbol{z}_{i,K}^o]$ with K classes to store confounder features. Direction-and-landmark keywords are extracted based on their part-of-speech tags [63]. We use the pre-trained RoBERTa [41] to obtain feature representations for each extracted token $\boldsymbol{f}_i^o$. Since the same word can have different features across sentences, we calculate the average feature for each keyword: $\boldsymbol{z}_{i,n}^o = \frac{1}{|z_{i,n}^o|}\sum_j \boldsymbol{f}_{i,n,j}^o$. Subsequently, the text content causal representation G_s is calculated as follows:

$$I = \text{RoBERTa}(\psi_t(\mathcal{I}) + \psi_t(\mathcal{P})),\ Z_k = \text{LN}(\phi_k(\mathcal{D}_I^O)) \quad (6)$$

$$G_s = \text{LN}[\phi_i(\mathcal{B}(I, Z_k))] \quad (7)$$

where $\psi(\cdot)$ and $\phi(\cdot)$ denote the learnable embedding layer and the full-connection layer, respectively. The absolute positional encoding $\mathcal{P}$ [60] is added to present the position information, and the layer normalization LN [5] is employed for stabilizing hidden states during training.

BACL in Vision Content. For each step, the panorama $\mathcal{V}$ is divided into 36 sub-images. Since existing VLN datasets primarily involve indoor room navigation, visual room references are treated as observable confounders. CLIP [56] is used to extract image features. Since room labels are not directly provided, we employ BLIP [32], a pre-trained VQA model to capture the room information for each image, by querying the model with a fixed prompt *"what kind of room is this?"* The average value of each room reference type is calculated, forming a visual room reference dictionary $\mathcal{D}_V^O = [\boldsymbol{z}_{v,1}^o, \boldsymbol{z}_{v,2}^o, ..., \boldsymbol{z}_{v,M}^o]$, where M is the number of room types. Additionally, the matrix $\gamma = [(\sin\theta_i, \cos\theta_i, \sin\eta_i, \cos\eta_i)_{i=1}^{36}]$ is used to present the direction of each image's shift relative to the agent, where θ and η denote the heading and elevation direction. If there are additional object features (for goal-oriented tasks), they are concatenated with image features. Subsequently, a 2-layer transformer encoder is used to capture spatial dependencies. The above process is formulated as follows:

$$V = \text{CLIP}(\mathcal{V}),\ Z_r = \text{LN}(\phi_r(\mathcal{D}_V^O)) \quad (8)$$

$$V_v = \text{LN}[\phi_v(\mathcal{B}(V, Z_r))] \quad (9)$$

$$Q_v = \text{Trans}(V_v + \psi_d(\gamma)) \quad (10)$$

4.2. Unobservable Causal Inference

Front-door Adjustment Causal Learning (FACL). In the previous section, we employed the back-door adjustment technique to handle bias caused by observable confounders. However, there are additional unobservable confounders that cannot be explicitly captured and modeled in advance. To address this, we introduce another technique - front-door adjustment [17]. As shown in Fig. 4(b), an additional mediator $\mathcal{M}$ is inserted between inputs and outcomes to construct the front door path $\mathcal{X} \rightarrow \mathcal{M} \rightarrow \mathcal{Y}$. In VLN, an attention-based model $P(\mathcal{Y}|\mathcal{X}) = \sum_m P(\mathcal{Y}|m)P(m|\mathcal{X})$ will select key regions $\mathcal{M}$ from inputs $\mathcal{X}$ for action prediction $\mathcal{Y}$. Therefore, the model inference can be represented by two parts: the feature selector $\mathcal{X} \rightarrow \mathcal{M}$ which selects suitable knowledge $\mathcal{M}$ from $\mathcal{X}$, and the action predictor

$\mathcal{M} \rightarrow \mathcal{Y}$ which exploits $\mathcal{M}$ to predict $\mathcal{Y}$. To eliminate spurious correlation brought by unobservable confounder $\mathcal{Z}$, we simultaneously deploy *do*-operator to $\mathcal{X}$ and $\mathcal{M}$:

$$P(\mathcal{Y}|do(\mathcal{X})) = \sum_m P(\mathcal{Y}|do(m))P(m|do(\mathcal{X})) \quad (11)$$

$$= \sum_{x'} P(x') \sum_m P(\mathcal{Y}|m, x')P(m|\mathcal{X}) \quad (12)$$

$$= \mathbb{E}_{x'}\mathbb{E}_{m|x}[P(\mathcal{Y}|x', m)] \quad (13)$$

where x' denotes potential input samples of the whole representation space, different from current inputs $\mathcal{X} = x$. We use the bold symbol $\boldsymbol{m}$ to denote the in-sampling features obtained by the feature extractor acting on the current input, and $\boldsymbol{x}'$ to mean the cross-sampling features randomly sampled by the K-means-based feature selector from the entire training samples. Based on the linear mapping model, Eq. (13) becomes $\mathbb{E}_{m|x}[\boldsymbol{m}] + \mathbb{E}_{x'}[\boldsymbol{x}']$. As it is intractable to get a closed-form solution of expectations involving the complex representation space, the estimation is achieved by the query mechanism. Two embedding functions [42, 73] are used to transmit input $\boldsymbol{x}$ into two query sets $\boldsymbol{g}_1 = q_1(\boldsymbol{x})$ and $\boldsymbol{g}_2 = q_2(\boldsymbol{x})$. Then, the front-door adjustment is approximated as follows:

$$\mathbb{E}_{x'}[\boldsymbol{x}'] \approx \sum_{x'} P(\boldsymbol{x}'|\boldsymbol{g}_1)\boldsymbol{x}' = \sum_i \frac{\exp(\boldsymbol{g}_1 \boldsymbol{x}_i'^T)}{\sum_j \exp(\boldsymbol{g}_1 \boldsymbol{x}_j'^T)} \boldsymbol{x}_i' \quad (14)$$

$$\mathbb{E}_{m|x}[\boldsymbol{m}] \approx \sum_m P(\boldsymbol{m}|\boldsymbol{g}_2)\boldsymbol{m} = \sum_i \frac{\exp(\boldsymbol{g}_2 \boldsymbol{m}_i^T)}{\sum_j \exp(\boldsymbol{g}_2 \boldsymbol{m}_j^T)} \boldsymbol{m}_i \quad (15)$$

$$\mathcal{F}(\boldsymbol{x}, \boldsymbol{x}') = \mathbb{E}_{x'}[\boldsymbol{x}'] + \mathbb{E}_{m|x}[\boldsymbol{m}] \quad (16)$$

The above process can be efficiently implemented using the multi-head attention [60], enabling seamless integration of causal adjustments into existing transformer-based frameworks with minimal modifications.

FACL in Text, Vision, and History. Considering the characteristics of VLN, we propose to eliminate unobservable confounders from three kinds of inputs in VLN, *i.e.*, vision, language, and history. First, following previous graph-based methods [9, 63], we construct the vision sequence $G_v = \{\texttt{[STOP]}, \texttt{[MEM]}, Q_v\}$ and the history sequence $G_h = \{\texttt{[STOP]}, \texttt{[MEM]}, \{\bar{Q}_t\}_{t=1}^T\}$ by adding the additional token for presenting stop and recurrent memory states, respectively. $\bar{Q}_t$ means the learned weight sum of panoramic features for the t-th step. To condense the lengthy sequence of features and generate global features $\boldsymbol{x}'$ for cross-sampling, we devise the CFP module (as described in Sec. 4.3) with the attentive pooling mechanism to construct confounder dictionaries for vision, history, and instruction, denoted as D_v^U, D_h^U, and D_s^U, respectively. Then the causality-enhanced features R_v, R_h and R_s are calculated based on Eq. (16):

$$R_v = \mathcal{F}(G_v, D_v^U), R_h = \mathcal{F}(G_h, D_h^U), R_s = \mathcal{F}(G_s, D_s^U) \quad (17)$$

Furthermore, we introduce an adaptive gate fusion (AGF) method to enhance the stability of learning by integrating causality-enhanced features with the original context features for each modality:

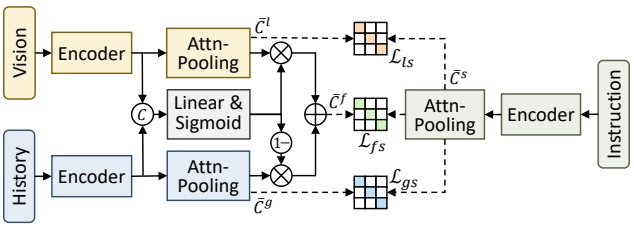

Figure 5. Illustration of the cross-modal feature pooling (CFP).

$$\omega_x = \delta(R_x W_x + G_x W_g + b) \quad (18)$$

$$U_x = \omega_x \odot R_x + (1 - \omega_x) \odot G_x \quad (19)$$

where δ and $\odot$ mean the Sigmoid function and element-wise multiplication. Suppose R_x and $G_x \in \mathbb{R}^{L_x \times d_h}$, then $W_{x/g} \in \mathbb{R}^{d_h \times 1}$ and $b \in \mathbb{R}^{L_x \times 1}$ are learnable parameters. Next, the cross-modal fused local features F_l and global features F_g are obtained by the cross-attention encoders $\mathcal{C}$ from METER [15]. The dynamic fusion $\mathcal{DF}$ [9] followed by Softmax $\mathcal{SF}$ is applied for action prediction:

$$F_l = \mathcal{C}(U_v, U_s, U_s),\ F_g = \mathcal{C}(U_h, U_s, U_s) \quad (20)$$

$$F_f = \mathcal{DF}(F_l, F_g),\ y_t = \mathcal{SF}(F_f) \quad (21)$$

The cross-entropy loss is used to optimize the network:

$$\mathcal{L}_{ce} = \sum_{t=1}^{T} -\log P(y_t^*|\mathcal{I}, \mathcal{V}_t, \mathcal{H}_{1:t-1}) \quad (22)$$

4.3. Cross-modal Feature Pooling

One challenge of implementing the front-door adjustment in VLN is constructing efficient dictionaries for global features from long sequences. This requires compressing sequential features of varying lengths into a unified feature space to represent each sample effectively. Let $H \in \mathbb{R}^{L \times d_h}$ be the sequential features, the following attentive pooling is used to effectively compress the sequence length:

$$A = \mathcal{T}(H),\ \alpha = \mathcal{SF}(AW_a),\ \bar{H} = \mathcal{T}(\alpha^T H) \quad (23)$$

where $\mathcal{T}$ denote the Tanh activation, $W_a \in \mathbb{R}^{d_h \times 1}$ is the learnable attention matrix, and $\bar{H} \in \mathbb{R}^{1 \times d_h}$. As shown in Fig. 5, for vision, history, local-global fusion, and text features, we use one transformer layer as the encoder followed by the attentive pooling to obtain flattened features $\bar{C}^l, \bar{C}^g, \bar{C}^f$, and $\bar{C}^s$, respectively. Then, we adopt contrastive learning [35, 57, 62] to optimize this cross-modal feature pooling (CFP) module, meanwhile improving semantic alignments for different modalities. The contrastive loss $\mathcal{L}_{ls}$ is constructed as:

$$\mathcal{L}_{ls} = -\frac{1}{2B}\sum_{j=1}^{B} \log \frac{\exp(\langle \bar{C}_j^l, \bar{C}_j^s \rangle / t)}{\sum_{k=1}^{B} \exp(\langle \bar{C}_j^l, \bar{C}_k^s \rangle / t)} - \frac{1}{2B}\sum_{k=1}^{B} \log \frac{\exp(\langle \bar{C}_k^l, \bar{C}_k^s \rangle / t)}{\sum_{j=1}^{B} \exp(\langle \bar{C}_j^l, \bar{C}_k^s \rangle / t)} \quad (24)$$

where B and t mean the batch size and temperature, respectively. Similarly, contrastive losses $\mathcal{L}_{gs}$ and $\mathcal{L}_{fs}$ are calculated by replacing $\bar{C}^l$ with $\bar{C}^g$ and $\bar{C}^f$. The overall CFP loss is the sum of these losses $\mathcal{L}_{\text{CFP}} = \mathcal{L}_{ls} + \mathcal{L}_{gs} + \mathcal{L}_{fs}$.

To enable the network more adaptive to characteristics for VLN and thus facilitate the building of the front-door confounder dictionaries for samples, we train the CFP module alongside other auxiliary tasks [9] during pre-training. Subsequently, the trained attentive pooling modules are used to extract global features for different modalities. In the fine-tuning stage, we employ the BACL and FACL with established dictionaries for intervention. The CFP offers dual benefits: it aligns different modalities more effectively during pre-training and provides a systematic approach for extracting coherent representations from sequence inputs.

4.4. Causal Learning Pipeline

As shown in Fig. 6, we summarize a causal learning pipeline that serves as a blueprint for similar learning-based methods. First, it begins with task formulation, where the specific task and its objectives are precisely defined. Next, the observable and unobservable confounders are explicitly assumed. Both back-door and front-door adjustment strategies are employed to tackle these confounders, either simultaneously or sequentially, contingent on task specifics. Then, the pipeline proceeds to model calculation and result prediction. Throughout network optimization, both network parameters and confounder features are continuously updated. Ultimately, this iterative process leads us to the development of a robust causal model capable of generating unbiased features, thereby advancing the generalizability of AI systems.

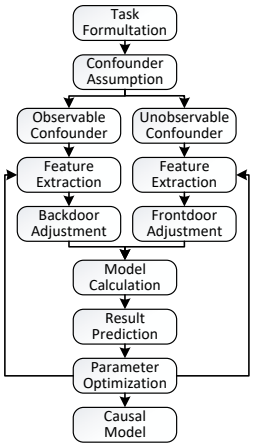

Figure 6. Pipeline of the causal learning.

5. Experiments

5.1. Experimental Settings

1) Datasets. We verify GOAT on two kinds of VLN benchmarks: fine-grained datasets (R2R [4] and RxR-English [29]), which provide long step-by-step navigation instructions, and goal-oriented datasets (REVERIE [53] and SOON [81]), which additionally requires for the target object. Formally, the datasets are partitioned into four splits: training, validation seen (sharing the same environments with the training set), validation unseen (having different environments from the training set), and test unseen sets (reported by the online leaderboard for fair comparison).

2) Evaluation Metrics. In R2R, key metrics include Navigation Error (NE), Success Rate (SR), Oracle SR (OSR), and SR Weighted by Path Length (SPL). RxR adds Normalized Dynamic Time Warping (nDTW) and SR Weighted by Dynamic Time Warping (sDTW). REVERIE and SOON introduce Remote Grounding Success Rate (RGS) and RGS Weighted by Path Length (RGSPL).

3) Implementation Details. Our model consists of 6 transformer layers for text, 2 for panorama, and 3 for cross-modal encoding. We use CLIP-B/16 [56] for image feature extraction and initialize network weights with METER [15]. In pre-training, MLM [28], SAP [8], and the proposed CFP are for R2R and RxR. OG [37] is added for REVERIE and SOON. EnvEdit [34] is employed for feature augmentation. The synthetic extended datasets [19, 62, 65] are used for R2R, REVERIE, and RxR, respectively. Pre-training is done on a single Tesla V100 GPU for a maximum of 300K iterations by the AdamW [43] optimizer, with a batch size of 48 and a learning rate of 5×10^{-5}. The numbers of classes of keywords and rooms are 74 and 50, and the temperature t is set to 1. In fine-tuning, the front-door dictionaries are randomly sampled from the K-Means clustering features. As the text transformer RoBERTa [41] is involved in end-to-end training, the textual keywords dictionary is also iteratively updated. Speaker models with the environmental dropout [58, 62, 64] are used to provide dynamic pseudo labels. We employ batch size 12 for R2R, REVERIE, and 5 for RxR and SOON, with a learning rate of 2×10^{-5} and a maximum of 100K iterations.

5.2. Comparisons with State-of-the-Arts

In Tab. 1, 2, 3, 4, we compare GOAT with the previous state-of-the-art (SoTA) methods on the R2R, REVERIE, RxR-English, and SOON datasets, respectively. On all these four datasets, our approach exhibits superior navigation performance, precise instruction-following alignment, and accurate object grounding across both seen and unseen environments. For instance, in R2R, GOAT achieves remarkable improvements in SPL compared to BEVBert [3], with relative increases of 7.41%, 6.45%, and 4.74% on three subsets. In REVERIE, GOAT shows substantial relative enhancements in RGSPL by 11.83%, 6.55%, and 20.87% on three subsets. In the challenging SOON and RxR tasks, GOAT also exhibits significant improvements in performance metrics, highlighting its robustness and superior generalization capabilities over previous methods.

5.3. Quantitative Analysis

1) Effect of Causal Inference. Fig. 7 verifies the impact of causal inference on GOAT across four diverse VLN datasets in unseen environments. "W/o intervention" signifies the

Method	Validation Seen				Validation Unseen				Test Unseen			
	SR↑	SPL↑	NE↓	OSR↑	SR↑	SPL↑	NE↓	OSR↑	SR↑	SPL↑	NE↓	OSR↑
HAMT [8]	76	72	2.51	82	66	61	3.29	73	65	60	3.93	72
DUET [9]	79	73	2.28	86	72	60	3.31	81	69	59	3.65	76
TD-STP [80]	77	73	2.34	83	70	63	3.22	76	67	61	3.73	72
GeoVLN [25]	79	76	2.22	–	68	63	3.35	–	65	61	3.95	–
DSRG [63]	81	76	2.23	88	73	62	3.00	81	72	61	3.33	78
GridMM [71]	–	–	–	–	75	64	2.83	–	73	62	3.35	–
GELA [10]	76	73	2.39	–	71	65	3.11	–	67	62	3.59	–
EnvEdit [34]	77	74	2.32	–	69	64	3.24	–	68	64	3.59	–
BEVBert [3]	81	74	2.17	88	75	64	2.81	84	73	62	3.13	**81**
GOAT (Ours)	**83.74**	**79.48**	**1.79**	**88.64**	**77.82**	**68.13**	**2.40**	**84.72**	**74.57**	**64.94**	**3.04**	80.35

Table 1. Comparison with other state-of-the-art methods on the R2R dataset [4]. '–': unavailable statistics.

Method	Validation Seen				Validation Unseen				Test Unseen			
	SR↑	SPL↑	RGS↑	RGSPL↑	SR↑	SPL↑	RGS↑	RGSPL↑	SR↑	SPL↑	RGS↑	RGSPL↑
HAMT [8]	43.29	40.19	27.20	25.18	32.95	30.20	18.92	17.28	30.40	26.67	14.88	13.08
HOP+ [54]	55.87	49.55	40.76	36.22	36.07	31.13	22.49	19.33	33.82	28.24	20.20	16.86
DUET [9]	71.75	63.94	57.41	51.14	46.98	33.73	32.15	23.03	52.51	36.06	31.88	22.06
DSRG [63]	75.69	68.09	61.07	54.72	47.83	34.02	32.69	23.37	54.04	37.09	32.49	22.18
GridMM [71]	–	–	–	–	51.37	36.47	34.57	24.56	53.13	36.60	34.87	23.45
BEVBert [3]	73.72	65.32	57.70	51.73	51.78	36.37	34.71	24.44	52.81	36.41	32.06	22.09
GOAT (Ours)	**78.64**	**71.40**	**63.74**	**57.85**	**53.37**	**36.70**	**38.43**	**26.09**	**57.72**	**40.53**	**38.32**	**26.70**

Table 2. Comparison with other state-of-the-art methods on the REVERIE dataset [53]. '–': unavailable statistics.

Method	Validation Seen				Validation Unseen			
	SR↑	SPL↑	nDTW↑	sDTW↑	SR↑	SPL↑	nDTW↑	sDTW↑
Syntax [31]	48.1	44.0	58.0	40.0	39.2	35.0	52.0	32.0
SOAT [45]	–	–	–	–	44.2	–	54.8	36.4
HOP+ [54]	53.6	47.9	59.0	43.0	45.7	38.4	52.0	36.0
FOAM [14]	–	–	–	–	42.8	38.7	54.1	35.6
ADAPT [35]	50.3	44.6	56.3	40.6	46.9	40.2	54.1	37.7
$MAR_{M\text{-}MP}$ [27]	–	–	–	–	50.2	–	60.3	43.9
VLN-PETL [55]	60.5	56.8	65.7	51.7	57.9	54.2	64.9	49.7
GOAT (Ours)	**74.1**	**68.1**	**71.0**	**61.4**	**68.2**	**61.7**	**67.1**	**56.6**

Table 3. Comparison on the RxR-English dataset [29].

Method	Validation Unseen				Test Unseen			
	OSR↑	SR↑	SPL↑	RGSPL↑	OSR↑	SR↑	SPL↑	RGSPL↑
GBE [81]	28.54	19.52	13.34	1.16	21.45	12.90	9.23	0.45
DUET [9]	50.91	36.28	22.58	3.75	43.00	33.44	21.42	4.17
GridMM [71]	53.39	37.46	24.81	3.91	48.02	36.27	21.25	4.15
GOAT (Ours)	**54.69**	**40.35**	**28.05**	**5.85**	**50.63**	**40.50**	**25.18**	**6.10**

Table 4. Comparison on the SOON dataset [81].

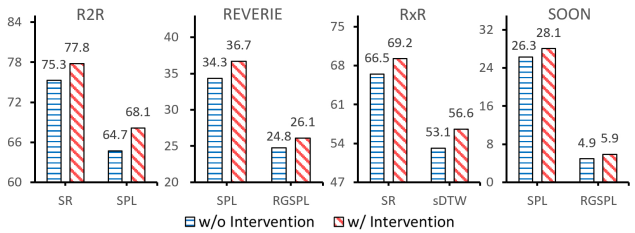

Figure 7. Effect of the intervention on various VLN datasets.

exclusion of the proposed BACL and FACL interventions across all modalities. On each of the datasets examined, the integration of causal inference leads to significant enhancements in the model's performance. This strongly demonstrates causal learning's considerable popularization potential in enhancing learning-based model generalization.

Id	BACL	FACL	SR↑	SPL↑	NE↓	OSR↑
1	✗	✗	75.27	64.69	2.72	83.14
2	✓	✗	76.37	67.13	2.60	83.99
3	✗	✓	76.50	66.92	2.53	84.21
4	✓	✓	**77.82**	**68.13**	**2.40**	**84.72**

Table 5. Effect of back-door and front-door adjustments.

2) Effect of BACL and FACL. Tab. 5 analyzes the effects of the proposed BACL and FACL on the R2R val-unseen subset. Compared to the baseline (#1), individual application of either BACL (#2) or FACL (#3) leads to performance improvements. Concurrent use of BACL and FACL (#4) leads to further performance enhancements. These findings underscore our assumption about the presence of both observable and unobservable confounders. Integrating both back-door and front-door adjustments is crucial to comprehensively addressing dataset biases, and enhancing the model's robustness and generalization.

3) Effect of CFP. In Tab. 6, we assess the efficacy of the proposed CFP on the R2R val-unseen subset. During pre-training (PT), incorporating CFP as an additional auxiliary task (CFP-P) enhances training performance, improving SR and SPL by 3.11% and 3.02% (#A1), respectively. In the fine-tuning (FT) stage, we compare the performance with and without the use of trained attention modules from CFP

Stage	Id	Method	SR↑	SPL↑	NE↓	OSR↑
PT	A0	w/o CFP-P	40.36	37.88	6.52	48.87
	A1	w/ CFP-P	43.47	40.90	6.06	53.13
FT	B0	A0 w/o CFP-F	75.56	65.90	2.63	82.42
	B1	A1 w/o CFP-F	76.63	66.17	2.63	84.67
	B2	A1 w/ CFP-F	**77.82**	**68.13**	**2.40**	**84.72**

Table 6. Effect of CFP in pre-training and fine-tuning.

Id	Text	Vision	SR↑	SPL↑	NE↓	OSR↑
1	Stats	Stats	75.22	64.78	2.71	83.65
2	Stats	Attn	76.59	65.39	2.56	**85.31**
3	**Attn**	**Stats**	**77.82**	**68.13**	**2.40**	84.72
4	Attn	Attn	75.95	65.83	2.64	83.91

Table 7. Effect of Statistic and Attention methods in BACL.

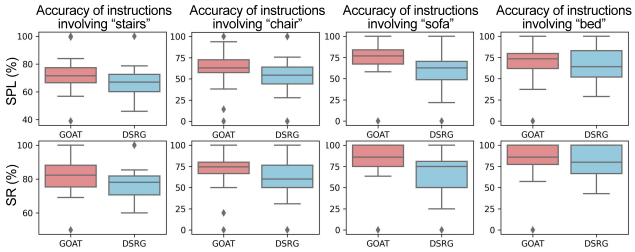

Figure 8. Comparison of the distribution of SR and SPL of instructions involving specific terms on the R2R val-unseen split.

to extract global features for front-door dictionaries (CFP-F). "W/o CFP-F" signifies the use of simple average pooling to compress features from the pre-trained model. #B2 shows that the CFP provides more reliable confounder representations for causal learning (SPL ↑ 1.96%).

4) Effect of Different BACL in Different Modalities. Tab. 7 investigates the effect of various combinations of statistic-based and attention-based methods for text and vision in BACL on the R2R val-unseen subset. The results indicate that employing the attention method for text and the statistic method for vision yields the best performance (#3). Intuitively, this can be explained by the structured nature of textual information and the involvement of RoBERTa's end-to-end training, enabling the attention method to effectively capture contextual nuances. Conversely, images lack explicit causality, and CLIP, the image extractor, isn't trained directly for efficiency reasons. Consequently, the statistic method ensures a stable causal learning process, preserving the integrity of vision-related features.

5.4. Qualitative Analysis

1) Bias Elimination Effect. In Fig. 8, the compactness of the boxes represents concentrated data distribution and reduced variability, while a median line closer to the center signifies even data distribution. It shows that GOAT obtains narrower boxes and more central midlines across diverse objects. This finding showcases that with the integration of causal intervention, GOAT significantly reduces prediction bias, thereby enhancing its generalization capability in previously unseen environments.

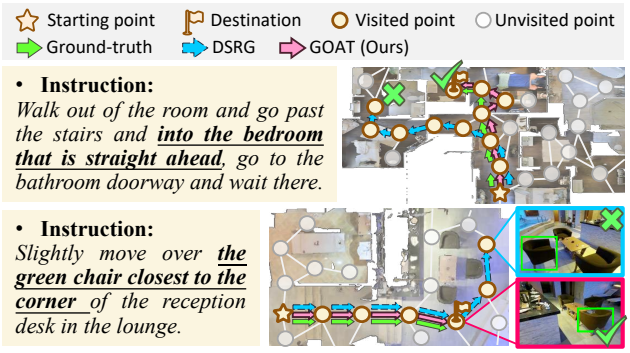

Figure 9. Predicted trajectories in unseen environments.

2) Visualized Trajectories. In Fig. 9, we visualize some predicted trajectories in unseen environments, comparing them with DSRG on R2R and REVERIE datasets. Notably, GOAT precisely captures directional cues like "straight ahead" and nuanced instructions like "closest to the corner", enabling accurate predictions. These instances highlight the intricate causal connections in VLN tasks, where specific instructions prompt corresponding actions. GOAT's enhanced causal inference capability enables it to generate more reasoned responses aligned with the provided instructions, underscoring the significance of robust causal inference in VLN systems. Please refer to our supplementary material for more detailed discussions and visualizations.

6. Conclusion

Our work presents GOAT, a novel approach that addresses the dataset bias in VLN from the perspective of causal learning. The back-door and front-door adjustment causal learning (BACL and FACL) mechanisms are proposed to adjust for observable and unobservable confounders, respectively. The cross-modal feature pooling (CFP) module is adopted to promote feature learning and extraction through contrastive learning. The practical causal learning pipeline is presented to illuminate other similar learning-based methods. Experiments on R2R, REVERIE, RxR, and SOON datasets show that GOAT can reasonably discover sequence visual-linguistic causal structures and significantly improve performance. Beyond VLN, the underlying confounder assumption and causal inference principles are generalizable to other similar fields.

Acknowledgments

This paper is supported by the National Natural Science Foundation of China under Grants (62233013, 62073245, 62173248). Shanghai Science and Technology Innovation Action Plan (22511104900).

References

[1] Ehsan Abbasnejad, Damien Teney, Amin Parvaneh, Javen Shi, and Anton van den Hengel. Counterfactual vision and language learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10044–10054, 2020. 2

[2] Dong An, Yuankai Qi, Yan Huang, Qi Wu, Liang Wang, and Tieniu Tan. Neighbor-view enhanced model for vision and language navigation. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 5101–5109, 2021. 1, 2

[3] Dong An, Yuankai Qi, Yangguang Li, Yan Huang, Liang Wang, Tieniu Tan, and Jing Shao. Bevbert: Multimodal map pre-training for language-guided navigation. *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023. 1, 6, 7

[4] 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 *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3674–3683, 2018. 1, 2, 6, 7, 14

[5] Jimmy Ba, Jamie Ryan Kiros, and Geoffrey E. Hinton. Layer normalization. *ArXiv*, abs/1607.06450, 2016. 4

[6] Pierre Baldi and Peter Sadowski. The dropout learning algorithm. *Artificial intelligence*, 210:78–122, 2014. 3

[7] 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. 2

[8] Shizhe Chen, Pierre-Louis Guhur, Cordelia Schmid, and Ivan Laptev. History aware multimodal transformer for vision-and-language navigation. *Advances in Neural Information Processing Systems*, 34, 2021. 1, 2, 6, 7

[9] Shizhe Chen, Pierre-Louis Guhur, Makarand Tapaswi, Cordelia Schmid, and Ivan Laptev. Think global, act local: Dual-scale graph transformer for vision-and-language navigation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16537–16547, 2022. 1, 2, 4, 5, 6, 7, 14

[10] Yibo Cui, Liang Xie, Yakun Zhang, Meishan Zhang, Ye Yan, and Erwei Yin. Grounded entity-landmark adaptive pre-training for vision-and-language navigation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12043–12053, 2023. 1, 7

[11] Ronghao Dang, Zhuofan Shi, Liuyi Wang, Zongtao He, Chengju Liu, and Qijun Chen. Unbiased directed object attention graph for object navigation. In *Proceedings of the 30th ACM International Conference on Multimedia*, page 3617–3627, New York, NY, USA, 2022. Association for Computing Machinery. 2

[12] Ronghao Dang, Lu Chen, Liuyi Wang, He Zongtao, Chengju Liu, and Qijun Chen. Multiple thinking achieving meta-ability decoupling for object navigation. In *International Conference on Machine Learning (ICML)*, 2023. 1

[13] Ronghao Dang, Liuyi Wang, Zongtao He, Shuai Su, Jiagui Tang, Chengju Liu, and Qijun Chen. Search for or navigate to? dual adaptive thinking for object navigation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8250–8259, 2023. 1

[14] Zi-Yi Dou and Nanyun Peng. Foam: A follower-aware speaker model for vision-and-language navigation. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4332–4340, 2022. 1, 7

[15] Zi-Yi Dou, Yichong Xu, Zhe Gan, Jianfeng Wang, Shuohang Wang, Lijuan Wang, Chenguang Zhu, Pengchuan Zhang, Lu Yuan, Nanyun Peng, et al. An empirical study of training end-to-end vision-and-language transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18166–18176, 2022. 5, 6

[16] 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. *Advances in Neural Information Processing Systems*, 31, 2018. 1, 2, 3

[17] Madelyn Glymour, Judea Pearl, and Nicholas P Jewell. *Causal inference in statistics: A primer*. John Wiley & Sons, 2016. 3, 4

[18] Pierre-Louis Guhur, Makarand Tapaswi, Shizhe Chen, Ivan Laptev, and Cordelia Schmid. Airbert: In-domain pretraining for vision-and-language navigation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1634–1643, 2021. 1, 2

[19] Weituo Hao, Chunyuan Li, Xiujun Li, Lawrence Carin, and Jianfeng Gao. Towards learning a generic agent for vision-and-language navigation via pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13137–13146, 2020. 2, 6

[20] Zongtao He, Liuyi Wang, Ronghao Dang, Shu Li, Qingqing Yan, Chengju Liu, and Qijun Chen. Learning depth representation from rgb-d videos by time-aware contrastive pre-training. *IEEE Transactions on Circuits and Systems for Video Technology*, pages 1–1, 2023. 1

[21] Zongtao He, Liuyi Wang, Shu Li, Qingqing Yan, Chengju Liu, and Qijun Chen. Mlanet: Multi-level attention network with sub-instruction for continuous vision-and-language navigation. *arXiv preprint arXiv:2303.01396*, 2023. 1, 2

[22] Yicong Hong, Qi Wu, Yuankai Qi, Cristian Rodriguez-Opazo, and Stephen Gould. Vln bert: A recurrent vision-and-language bert for navigation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1643–1653, 2021. 2

[23] Yicong Hong, Yang Zhou, Ruiyi Zhang, Franck Dernoncourt, Trung Bui, Stephen Gould, and Hao Tan. Learning navigational visual representations with semantic map supervision. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3055–3067, 2023. 1

[24] Ronghang Hu, Daniel Fried, Anna Rohrbach, Dan Klein, Trevor Darrell, and Kate Saenko. Are you looking? grounding to multiple modalities in vision-and-language navigation.

In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6551–6557, 2019. 1

[25] Jingyang Huo, Qiang Sun, Boyan Jiang, Haitao Lin, and Yanwei Fu. Geovln: Learning geometry-enhanced visual representation with slot attention for vision-and-language navigation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23212–23221, 2023. 7

[26] Vihan Jain, Gabriel Magalhaes, Alexander Ku, Ashish Vaswani, Eugene Ie, and Jason Baldridge. Stay on the path: Instruction fidelity in vision-and-language navigation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1862–1872, 2019. 1

[27] Aishwarya Kamath, Peter Anderson, Su Wang, Jing Yu Koh, Alexander Ku, Austin Waters, Yinfei Yang, Jason Baldridge, and Zarana Parekh. A new path: Scaling vision-and-language navigation with synthetic instructions and imitation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10813–10823, 2023. 2, 7

[28] Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*, pages 4171–4186, 2019. 6

[29] Alexander Ku, Peter Anderson, Roma Patel, Eugene Ie, and Jason Baldridge. Room-across-room: Multilingual vision-and-language navigation with dense spatiotemporal grounding. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4392–4412, 2020. 1, 2, 6, 7, 14

[30] Jialu Li and Mohit Bansal. Improving vision-and-language navigation by generating future-view image semantics. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10803–10812, 2023. 1

[31] Jialu Li, Hao Tan, and Mohit Bansal. Improving cross-modal alignment in vision language navigation via syntactic information. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1041–1050, 2021. 7

[32] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning*, pages 12888–12900. PMLR, 2022. 4

[33] Jiangtong Li, Li Niu, and Liqing Zhang. From representation to reasoning: Towards both evidence and commonsense reasoning for video question-answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21273–21282, 2022. 2

[34] Jialu Li, Hao Tan, and Mohit Bansal. Envedit: Environment editing for vision-and-language navigation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15407–15417, 2022. 1, 2, 6, 7

[35] Bingqian Lin, Yi Zhu, Zicong Chen, Xiwen Liang, Jianzhuang Liu, and Xiaodan Liang. Adapt: Vision-language navigation with modality-aligned action prompts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15396–15406, 2022. 1, 5, 7

[36] Kunyang Lin, Peihao Chen, Diwei Huang, Thomas H Li, Mingkui Tan, and Chuang Gan. Learning vision-and-language navigation from youtube videos. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8317–8326, 2023. 1, 2

[37] Xiangru Lin, Guanbin Li, and Yizhou Yu. Scene-intuitive agent for remote embodied visual grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7036–7045, 2021. 6

[38] Bing Liu, Dong Wang, Xu Yang, Yong Zhou, Rui Yao, Zhiwen Shao, and Jiaqi Zhao. Show, deconfound and tell: Image captioning with causal inference. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18041–18050, 2022. 2, 3

[39] Chong Liu, Fengda Zhu, Xiaojun Chang, Xiaodan Liang, Zongyuan Ge, and Yi-Dong Shen. Vision-language navigation with random environmental mixup. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1644–1654, 2021. 1, 2

[40] Rui Liu, Xiaohan Wang, Wenguan Wang, and Yi Yang. Bird's-eye-view scene graph for vision-language navigation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10968–10980, 2023. 1

[41] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019. 4, 6

[42] Yang Liu, Guanbin Li, and Liang Lin. Cross-modal causal relational reasoning for event-level visual question answering. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023. 2, 3, 5

[43] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2018. 6

[44] Arjun Majumdar, Ayush Shrivastava, Stefan Lee, Peter Anderson, Devi Parikh, and Dhruv Batra. Improving vision-and-language navigation with image-text pairs from the web. In *European Conference on Computer Vision*, pages 259–274. Springer, 2020. 1, 2

[45] Abhinav Moudgil, Arjun Majumdar, Harsh Agrawal, Stefan Lee, and Dhruv Batra. Soat: A scene-and object-aware transformer for vision-and-language navigation. *Advances in Neural Information Processing Systems*, 34:7357–7367, 2021. 7

[46] Laksh Nanwani, Anmol Agarwal, Kanishk Jain, Raghav Prabhakar, Aaron Monis, Aditya Mathur, Krishna Murthy, Abdul Hafez, Vineet Gandhi, and K Madhava Krishna. Instance-level semantic maps for vision language navigation. *arXiv preprint arXiv:2305.12363*, 2023. 1

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

on Natural Language Processing (EMNLP-IJCNLP), pages 684–695, 2019. 16

[48] Siqi Nie, Meng Zheng, and Qiang Ji. The deep regression bayesian network and its applications: Probabilistic deep learning for computer vision. *IEEE Signal Processing Magazine*, 35(1):101–111, 2018. 3

[49] Yulei Niu, Kaihua Tang, Hanwang Zhang, Zhiwu Lu, Xian-Sheng Hua, and Ji-Rong Wen. Counterfactual vqa: A cause-effect look at language bias. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12700–12710, 2021. 2

[50] Amin Parvaneh, Ehsan Abbasnejad, Damien Teney, Javen Qinfeng Shi, and Anton Van den Hengel. Counterfactual vision-and-language navigation: Unravelling the unseen. *Advances in Neural Information Processing Systems*, 33:5296–5307, 2020. 2

[51] Judea Pearl. *Causality*. Cambridge university press, 2009. 1

[52] Judea Pearl and Dana Mackenzie. *The book of why: the new science of cause and effect*. Basic books, 2018. 2, 3, 13

[53] Yuankai Qi, Qi Wu, Peter Anderson, Xin Wang, William Yang Wang, Chunhua Shen, and Anton van den Hengel. Reverie: Remote embodied visual referring expression in real indoor environments. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9982–9991, 2020. 2, 3, 6, 7, 14

[54] Yanyuan Qiao, Yuankai Qi, Yicong Hong, Zheng Yu, Peng Wang, and Qi Wu. Hop+: History-enhanced and order-aware pre-training for vision-and-language navigation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023. 7

[55] Yanyuan Qiao, Zheng Yu, and Qi Wu. Vln-petl: Parameter-efficient transfer learning for vision-and-language navigation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15443–15452, 2023. 7

[56] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021. 2, 4, 6

[57] Sheng Shen, Liunian Harold Li, Hao Tan, Mohit Bansal, Anna Rohrbach, Kai-Wei Chang, Zhewei Yao, and Kurt Keutzer. How much can clip benefit vision-and-language tasks? In *International Conference on Learning Representations*. 5

[58] Hao Tan, Licheng Yu, and Mohit Bansal. Learning to navigate unseen environments: Back translation with environmental dropout. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2610–2621, 2019. 2, 6

[59] Jesse Thomason, Michael Murray, Maya Cakmak, and Luke Zettlemoyer. Vision-and-dialog navigation. In *Conference on Robot Learning*, pages 394–406. PMLR, 2020. 16

[60] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017. 4, 5

[61] Hanqing Wang, Wei Liang, Jianbing Shen, Luc Van Gool, and Wenguan Wang. 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*, pages 15471–15481, 2022. 1

[62] Liuyi Wang, Zongtao He, Ronghao Dang, Huiyi Chen, Chengju Liu, and Qijun Chen. Res-sts: Referring expression speaker via self-training with scorer for goal-oriented vision-language navigation. *IEEE Transactions on Circuits and Systems for Video Technology*, 2023. 1, 2, 5, 6

[63] Liuyi Wang, Zongtao He, jiagui Tang, Ronghao Dang, naijia Wang, Chengju Liu, and Qijun Chen. A dual semantic-aware recurrent global-adaptive network for vision-and-language navigation. In *International Joint Conferences on Artificial Intelligence (IJCAI)*, 2023. 1, 2, 4, 5, 7

[64] Liuyi Wang, Chengju Liu, Zongtao He, Shu Li, Qingqing Yan, Huiyi Chen, and Qijun Chen. Pasts: Progress-aware spatio-temporal transformer speaker for vision-and-language navigation. *arXiv preprint arXiv:2305.11918*, 2023. 1, 2, 6

[65] Su Wang, Ceslee Montgomery, Jordi Orbay, Vighnesh Birodkar, Aleksandra Faust, Izzeddin Gur, Natasha Jaques, Austin Waters, Jason Baldridge, and Peter Anderson. Less is more: Generating grounded navigation instructions from landmarks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15428–15438, 2022. 1, 6

[66] Tan Wang, Jianqiang Huang, Hanwang Zhang, and Qianru Sun. Visual commonsense r-cnn. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10760–10770, 2020. 2, 3

[67] Tan Wang, Chang Zhou, Qianru Sun, and Hanwang Zhang. Causal attention for unbiased visual recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3091–3100, 2021. 2

[68] Xin Wang, Qiuyuan Huang, Asli Celikyilmaz, Jianfeng Gao, Dinghan Shen, Yuan-Fang Wang, William Wang, and Lei Zhang. Vision-language navigation policy learning and adaptation. *IEEE transactions on pattern analysis and machine intelligence*, 2020. 2

[69] Yuqing Wang, Xiangxian Li, Zhuang Qi, Jingyu Li, Xuelong Li, Xiangxu Meng, and Lei Meng. Meta-causal feature learning for out-of-distribution generalization. In *Computer Vision–ECCV 2022 Workshops: Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part VI*, pages 530–545. Springer, 2023. 2

[70] Zun Wang, Jialu Li, Yicong Hong, Yi Wang, Qi Wu, Mohit Bansal, Stephen Gould, Hao Tan, and Yu Qiao. Scaling data generation in vision-and-language navigation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12009–12020, 2023. 1

[71] Zihan Wang, Xiangyang Li, Jiahao Yang, Yeqi Liu, and Shuqiang Jiang. Gridmm: Grid memory map for vision-and-language navigation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15625–15636, 2023. 7

[72] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models, 2023. 16

[73] Xu Yang, Hanwang Zhang, and Jianfei Cai. Deconfounded image captioning: A causal retrospect. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021. 1, 2, 3, 5

[74] Xu Yang, Hanwang Zhang, Guojun Qi, and Jianfei Cai. Causal attention for vision-language tasks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9847–9857, 2021. 2

[75] Zhongqi Yue, Hanwang Zhang, Qianru Sun, and Xian-Sheng Hua. Interventional few-shot learning. *Advances in neural information processing systems*, 33:2734–2746, 2020. 2

[76] Hua Zhang, Liqiang Xiao, Xiaochun Cao, and Hassan Foroosh. Multiple adverse weather conditions adaptation for object detection via causal intervention. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–1, 2022. 2

[77] Jiazhao Zhang, Liu Dai, Fanpeng Meng, Qingnan Fan, Xuelin Chen, Kai Xu, and He Wang. 3d-aware object goal navigation via simultaneous exploration and identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6672–6682, 2023. 1

[78] Shengyu Zhang, Tan Jiang, Tan Wang, Kun Kuang, Zhou Zhao, Jianke Zhu, Jin Yu, Hongxia Yang, and Fei Wu. Devlbert: Learning deconfounded visio-linguistic representations. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 4373–4382, 2020. 2, 3

[79] Yubo Zhang, Hao Tan, and Mohit Bansal. Diagnosing the environment bias in vision-and-language navigation. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence*, 2021. 1

[80] Yusheng Zhao, Jinyu Chen, Chen Gao, Wenguan Wang, Lirong Yang, Haibing Ren, Huaxia Xia, and Si Liu. Target-driven structured transformer planner for vision-language navigation. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 4194–4203, 2022. 7

[81] Fengda Zhu, Xiwen Liang, Yi Zhu, Qizhi Yu, Xiaojun Chang, and Xiaodan Liang. Soon: Scenario oriented object navigation with graph-based exploration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12689–12699, 2021. 2, 3, 6, 7, 14

[82] Wanrong Zhu, Yuankai Qi, Pradyumna Narayana, Kazoo Sone, Sugato Basu, Xin Wang, Qi Wu, Miguel Eckstein, and William Yang Wang. Diagnosing vision-and-language navigation: What really matters. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5981–5993, 2022. 1