

# DELAN: Dual-Level Alignment for Vision-and-Language Navigation by Cross-Modal Contrastive Learning

Mengfei Du<sup>1†</sup>, Binhao Wu<sup>1†</sup>, Jiwen Zhang<sup>1</sup>, Zhihao Fan<sup>2</sup>,  
Zejun Li<sup>1</sup>, Ruipu Luo<sup>1</sup>, Xuanjing Huang<sup>3</sup>, Zhongyu Wei<sup>1\*</sup>

<sup>1</sup>School of Data Science, Fudan University, China

<sup>2</sup>Alibaba Group, China

<sup>3</sup>School of Computer Science, Fudan University, China

{mfdu22, bhwu22, jiwenzhang21, rpluo21}@m.fudan.edu.cn

{zejunli20, xjhuang, zywei}@fudan.edu.cn

## Abstract

Vision-and-Language navigation (VLN) requires an agent to navigate in unseen environment by following natural language instruction. For task completion, the agent needs to align and integrate various navigation modalities, including instruction, observation and navigation history. Existing works primarily concentrate on cross-modal attention at the fusion stage to achieve this objective. Nevertheless, modality features generated by disparate uni-encoders reside in their own spaces, leading to a decline in the quality of cross-modal fusion and decision. To address this problem, we propose a **Dual-level Alignment (DELAN)** framework by cross-modal contrastive learning. This framework is designed to align various navigation-related modalities before fusion, thereby enhancing cross-modal interaction and action decision-making. Specifically, we divide the pre-fusion alignment into dual levels: instruction-history level and landmark-observation level according to their semantic correlations. We also reconstruct a dual-level instruction for adaptation to the dual-level alignment. As the training signals for pre-fusion alignment are extremely limited, self-supervised contrastive learning strategies are employed to enforce the matching between different modalities. Our approach seamlessly integrates with the majority of existing models, resulting in improved navigation performance on various VLN benchmarks, including R2R, R4R, RxR and CVDN.

**Keywords:** Vision-and-Language Navigation, Cross-modal Alignment, Cross-modal Contrastive Learning

## 1. Introduction

Vision-and-Language navigation (VLN) has attracted increasing interest from natural language processing, computer vision and robotics communities due to its potential of applications such as embodied robots (Gu et al., 2022). The VLN task requires the agent to navigate to a predetermined destination, guided by natural language instruction and real-time visual observations. As depicted in Figure 1, a comprehensive navigation instruction typically comprises a sequence of action, direction and landmark clues. The agent needs to proficiently model its historical experiences and accurately relate the landmark clues to the environmental observations in order to successfully navigate.

Similar to traditional Vision-Language task, the general model architecture of VLN agent adopts a multi-modal late-fusion paradigm (Du et al., 2023). Under this paradigm, individual modalities are encoded by their respective encoders and then passed through a fusion module, as illustrated in Figure 2. In Vision-Language task, aligning unimodal representations before fusion has been extensively validated as it contributes to cross-modal modeling ability in the fusion stage and seman-

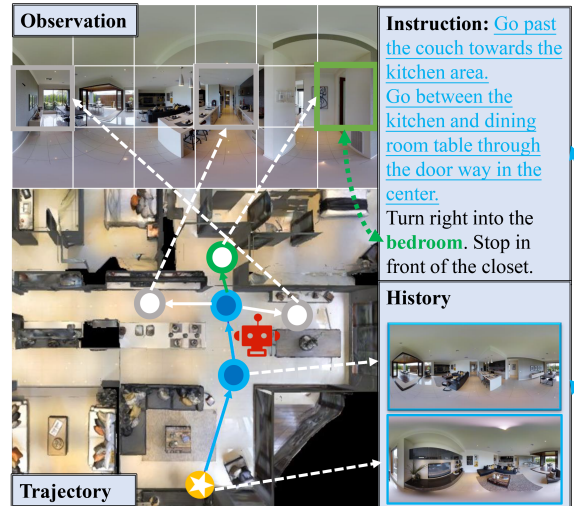


Figure 1: An Example. The VLN agent needs to align the historical trajectory to the corresponding instructions (e.g. "Go past the couch towards. . .") during navigation. At each time step, it selects one of the candidate viewpoints by evaluating the current observation in different directions and landmarks from the instructions (e.g. "bedroom").

tic comprehension ability of unimodal encoders (Li et al., 2021, 2022b). However, pre-fusion alignment has not yet been explored in the context of VLN

\* Corresponding author.

† Equal contribution.

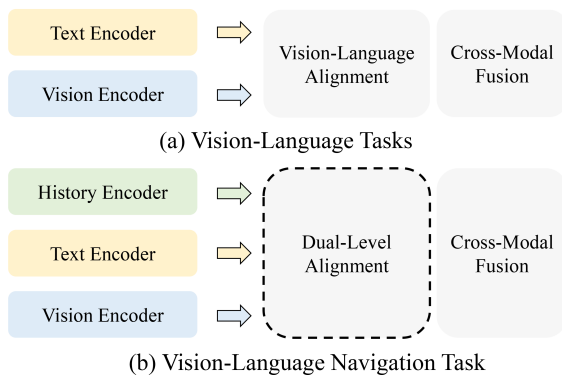


Figure 2: Comparison of pipelines for traditional VL and VLN tasks. (a) In VL tasks, vision-language pre-fusion alignment is often employed to encourage modality interaction. (b) However in VLN filed, there is a blank in pre-fusion alignment, which is solved by our dual-level alignment framework.

tasks, resulting in non-negligible disparities among correlated modalities before fusion.

Therefore, we propose to align corresponding unimodal representations and bridge the modality gaps prior to the cross-modal fusion stage in VLN task. This idea is challenging for two reasons. Firstly, there are limited training signals available for aligning unimodal representations. Training signals about navigation, such as navigation progress and relative spatial position, are used for supervising post-fusion representations in previous works (Ma et al., 2019a; Chen et al., 2021). Secondly, VLN is a sequential decision process distinct from traditional Vision-Language tasks (Hong et al., 2021). The navigation history condenses visited panorama and performed actions to retain past information, while observation at each time step contains detailed landmark semantics corresponding to the instruction. Consequently, how to effectively leverage the different characteristics of navigation history and real-time observations to enhance the cross-modal alignment remains an open question.

To address these problems, we propose a model-agnostic **Dual-level Alignment** (DELAN) framework that utilizes cross-modal contrastive learning. As the training signals are limited, we propose the self-supervised contrastive learning approach, relying on the contrast between positive and negative pairs within a batch as the training signal. For the diverse characteristics of unimodal representations, we partition the cross-modal pre-fusion alignment into two separate levels: instruction-history level alignment and landmark-observation level alignment, according to their strong semantic correlations. Specifically, history modality records visual experiences along the described navigation trajectory. The landmarks, such as “couch” and “closet”,

appearing in the real-time observations constitute the textual foundation of language instructions. Furthermore, we restructure the original instruction into a dual-level instruction-landmark format to explicitly formulate the dual-level semantic correspondences with history and observation. For instruction-history alignment, we employ a contrastive loss on the history tokens and the instruction part of dual-level instruction across both global and local representations. For landmark-observation alignment, we use a contrastive loss on the observations and the landmark part of dual-level instruction at each time step across only local representations.

We summarize our contributions as follows: (1) We are the first to introduce a cross-modal contrastive learning framework to improve the pre-fusion alignment in VLN task; (2) Inspired by the characteristics of different VLN modalities, we propose to conduct the dual-level alignment before fusion, i.e. instruction-history level alignment and landmark-observation level alignment, and develop self-supervised learning strategies respectively. (3) We validate our framework across various VLN benchmarks, demonstrating the effectiveness and consistency of our DELAN framework.<sup>1</sup>

## 2. Related Work

### 2.1. Vision-and-Language Navigation

In recent years, training an instruction-following navigator has attracted growing research interest. Earlier studies mainly employ a sequence-to-sequence LSTM framework in VLN tasks (Fried et al., 2018; Tan et al., 2019). Owing to the success of transformer (Vaswani et al., 2017b) in Vision-Language tasks (Tan and Bansal, 2019; Lu et al., 2019; Li et al., 2020), transformer-based models are widespread in the VLN field. VLN<sub>o</sub>BERT (Hong et al., 2021) devises a recurrent unit in the cross-modal transformer to model the navigation history. HAMT (Chen et al., 2021) proposes to encode all past observations and actions with hierarchical transformer. DUET (Chen et al., 2022) utilizes a graph transformer to encode global history for efficient planning. Besides the improvement of model architecture, various other methods are proposed to enhance the navigation performance, including training strategies (Wang et al., 2019; Zhang et al., 2021; Liang et al., 2022) and auxiliary tasks (Ma et al., 2019b; Zhu et al., 2020).

Despite the progress above, aligning navigation modalities remains a formidable challenge. OAAM (Qi et al., 2020) and RelGraph (Hong et al., 2020a) utilize separate soft-attention modules to implicitly align visual representations with different at-

<sup>1</sup>Code is available at <https://github.com/mengfeidu/DELAN>

tribute words. FGR2R (Hong et al., 2020b) grounds the observation to a certain sub-instruction using human annotations. DDL (Cheng et al., 2022) provides token-level supervision by labeling landmark and action words relating to observation at each time step. However, these approaches only focus on the cross-modal fusion module, neglecting the modality gaps among unimodal representations that are commonly appeared in vision-language tasks (Li et al., 2021). In contrast, our work presents a self-supervised framework to align these modalities before fusion stage by contrastive learning to promote cross-modal reasoning.

## 2.2. Cross-modal Contrastive Learning

Cross-modal contrastive learning has proven its effectiveness in modeling vision-language relationships and has gained prominence as a popular strategy in VL tasks. CLIP (Radford et al., 2021) firstly employs cross-modal contrastive learning on large-scale image-text pairs. FILIP (Yao et al., 2021) utilizes token-wise maximum similarity to achieve the fine-grained alignment between visual and textual tokens. TACo (Yang et al., 2021) proposes a token-aware contrastive loss and a cascade sampling method for aligning video and text. X-CLIP (Ma et al., 2022) proposes attention method to realize multi-grained contrast between text and video. CITL (Liang et al., 2022) first introduces the contrastive learning into VLN domain but only considers the unimodal contrast. Inspired by these works, we propose to apply contrastive learning to bridge the modality gaps. Furthermore, to handle the sparse training signals for pre-fusion alignment, we apply self-supervised learning strategies to realize our dual-level alignment framework.

## 3. Preliminaries

### 3.1. Problem Formulation

Given the instruction  $\mathcal{I} = \{w_1, w_2, \dots, w_m\}$ , the agent receives a panorama observation  $\mathcal{O}_t$  of its surrounding environment at each step  $t$ . The panorama observation consists of  $k$  single view images with corresponding relative angles, i.e.  $\mathcal{O}_t \triangleq ([v_1^o; a_1^o], \dots, [v_k^o; a_k^o])$ , where  $v_i^o$  is the visual feature of the  $i$ -th view and  $a_i^o$  is the relative angle with respect to the current view (subscript  $t$  is omitted for simplicity). At each viewpoint, the agent has  $n_t$  navigable candidate viewpoints  $\mathcal{O}_t^c \triangleq ([v_1^c; a_1^c], \dots, [v_{n_t}^c; a_{n_t}^c])$ . All observations  $\mathcal{O}_i$  and performed actions  $a_i^h$  before time step  $t$  form the history  $\mathcal{H}_t \triangleq ([\mathcal{O}_1; a_1^h], \dots, [\mathcal{O}_{t-1}; a_{t-1}^h])$ , where action  $a_i^h$  denotes the turned angles at step  $i$ . The navigation is finished when the agent makes **STOP** decision. The task is considered accomplished when the agent stops near the target location ( $<3m$ ).

### 3.2. Base Agent Model

Current VLN agents (Chen et al., 2021, 2022) mainly adopts multi-modal late-fusion learning paradigm (Du et al., 2023). To encode each modality, they usually employ standard BERT (Devlin et al., 2018) for textual instructions and vision transformer (ViT) (Dosovitskiy et al., 2021) for visual observations. As for history modality, they encode historical observations and performed actions during navigation, enabling agents to remember the trajectory. Specially, graph-based methods (Chen et al., 2022) maintain a topological map to record the historical experiences. Above three unimodal encoding processes can be formulated as follows:

$$\begin{aligned} \{x_{cls}, x_1, x_2, \dots, x_m\} &= \text{TextEnc}([\text{CLS}]; \mathcal{I}) \\ \{o_{t,1}, o_{t,2}, \dots, o_{t,k}\} &= \text{VisionEnc}(\mathcal{O}_t) \\ \{h_1, h_2, \dots, h_{t-1}\} &= \text{HistoryEnc}(\mathcal{H}_t) \end{aligned} \quad (1)$$

where  $x_i, o_{t,i}, h_i \in \mathbb{R}^d$ . These unimodal embeddings are further fed to a cross-modal transformer to predict the action probability over the candidate viewpoints  $\mathcal{O}_t^c$ . The candidate viewpoint with largest probability will be chosen as next move location.

### 3.3. Contrastive Learning Base

**Similarity reduce function.** Given a similarity vector/matrix  $M \in \mathbb{R}^{n \times m}$  of two instances from distinct modalities, we need to reduce it to scalar value for later contrastive loss calculation. Attention mechanism is introduced in the aggregation process, allowing each element to attend to multiple contrasting elements and automatically balance their influences (Ma et al., 2022). Concretely, we first perform attention operations on row and column to reduce values in each dimension.

$$\begin{aligned} M_r &= \sum_{i=1}^n \frac{\exp(M_{(i,*)})}{\sum_{j=1}^m \exp(M_{(j,*)})} \odot M_{(i,*)} \\ M_c &= \sum_{i=1}^m \frac{\exp(M_{(*,i)})}{\sum_{j=1}^n \exp(M_{(*,j)})} \odot M_{(*,i)} \end{aligned} \quad (2)$$

where  $*$  represents all elements in the dimension,  $\odot$  represents element-wise product,  $M_r \in \mathbb{R}^{1 \times m}$  and  $M_c \in \mathbb{R}^{n \times 1}$ . Likewise, we second perform attention operations on column and row to get a scale value respectively.

$$\begin{aligned} M'_r &= \sum_{i=1}^m \frac{\exp(M_{r(1,i)})}{\sum_{j=1}^m \exp(M_{r(1,j)})} M_{r(1,i)} \\ M'_c &= \sum_{i=1}^n \frac{\exp(M_{c(i,1)})}{\sum_{j=1}^n \exp(M_{c(j,1)})} M_{c(i,1)} \end{aligned} \quad (3)$$

where  $M'_r, M'_c \in \mathbb{R}^1$ . Finally, we average two scale values to acquire similarity score of two instances

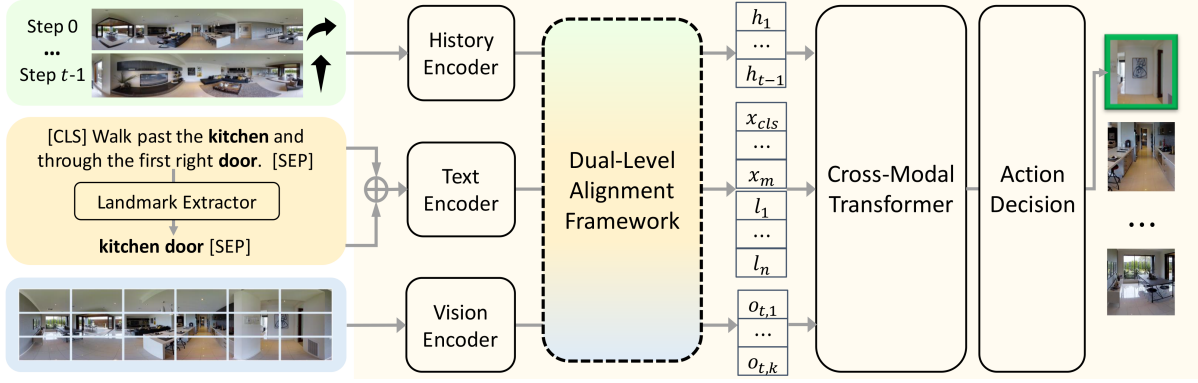


Figure 3: The pipeline of VLN model enhanced with DELAN framework. Landmark words are extracted from the original instructions, and three unimodal encoders encode four types of tokens: instructions, landmarks, histories, and observations. Then, the alignments are employed at pre-fusion stage to enhance cross-modal interaction and action decision.

and define above process as reduce function  $R$ .

$$R(M) = (M'_r + M'_c)/2 \quad (4)$$

**General contrastive loss.** Under a general formulation of cross-modal contrastive learning (Radford et al., 2021), we consider  $B$  pairs  $\{\mathbf{t}^{(i)}, \mathbf{v}^{(i)}\}_{i=1}^B$  from two modalities  $\mathbf{t}$  and  $\mathbf{v}$  in each training batch. With similarity scores  $\{sim(\mathbf{t}^{(i)}, \mathbf{v}^{(j)})\}_{i,j=1}^B \in \mathbb{R}$  among instances  $\mathbf{t}^{(i)}$  and  $\mathbf{v}^{(j)}$ , we can compute both  $\mathbf{t}$ -to- $\mathbf{v}$  and  $\mathbf{v}$ -to- $\mathbf{t}$  contrastive losses. The total contrastive loss can be summed from above losses and defined as contrastive loss function  $L$  as:

$$\begin{aligned} loss(\mathbf{t}, \mathbf{v}) &= -\frac{1}{B} \sum_{i=1}^B \log \frac{\exp(sim(\mathbf{t}^{(i)}, \mathbf{v}^{(i)})/\tau)}{\sum_{j=1}^B \exp(sim(\mathbf{t}^{(i)}, \mathbf{v}^{(j)})/\tau)} \\ loss(\mathbf{v}, \mathbf{t}) &= -\frac{1}{B} \sum_{i=1}^B \log \frac{\exp(sim(\mathbf{v}^{(i)}, \mathbf{t}^{(i)})/\tau)}{\sum_{j=1}^B \exp(sim(\mathbf{v}^{(i)}, \mathbf{t}^{(j)})/\tau)} \\ L(\mathbf{t}, \mathbf{v}) &= loss(\mathbf{t}, \mathbf{v}) + loss(\mathbf{v}, \mathbf{t}) \end{aligned} \quad (5)$$

where  $\tau$  is temperature parameter.

## 4. Methods

In this section, we elaborate each component of our dual-level alignment (DELAN) framework. We explain the dual-level instruction construction in section 4.1. Then we introduce the instruction-history level alignment in section 4.2, and the landmark-observation level alignment in section 4.3, both by self-supervised contrastive learning. Finally, we present the complete training strategy in section 4.4. The pipeline of our framework is illustrated as Figure 3 and Figure 4.

### 4.1. Dual-level Instruction Construction

A navigation instruction (e.g. "Walk to the right past the TV.") comprises a series of action, orientation and landmark words. Since the navigation history has strong semantic correlation with the entire instruction, while the real-time observations are closely related to the landmark words in instructions, it is significant to explicitly represent such dual-level semantic correspondences. Therefore, an additional landmark part is introduced for alignment with observation while the original instruction part is reserved for alignment with history.

Specifically, we concatenate original instruction with the extracted landmark words to form a dual-level instruction. We employ a standard pre-trained BERT model (Vaswani et al., 2017a) fine-tuned on part-of-speech task to extract all nouns in the instruction as landmark words. The instruction level text guides the overall navigation and the landmark level text helps focus on key details in observation. After this we can drive the instruction part align to history modality and the landmark part align to observation modality respectively.

Formally, we extract  $n$  landmark words  $\mathcal{I}^l = \{w_1^l, w_2^l, \dots, w_n^l\}$  from original instructions. These landmark words are appended to the original instructions, subsequently fed into the text encoder to get textual embeddings, i.e.

$$\{x_{cls}, X, L\} = \text{TextEnc}([\text{CLS}]; \mathcal{I}; \mathcal{I}^l) \quad (6)$$

where  $X = (x_1, x_2, \dots, x_m)$  is the instruction word matrix,  $L = (l_1, l_2, \dots, l_n)$  is the landmark word matrix. The representation of navigation history and real-time observations are computed based on equation 1, resulting in history embeddings matrix  $H = (h_1, h_2, \dots, h_T)$  and observation embeddings matrix  $O_t = (o_{t,1}, o_{t,2}, \dots, o_{t,k})$  at time step  $t$ . Note that  $T$  is the total number of navigation steps.

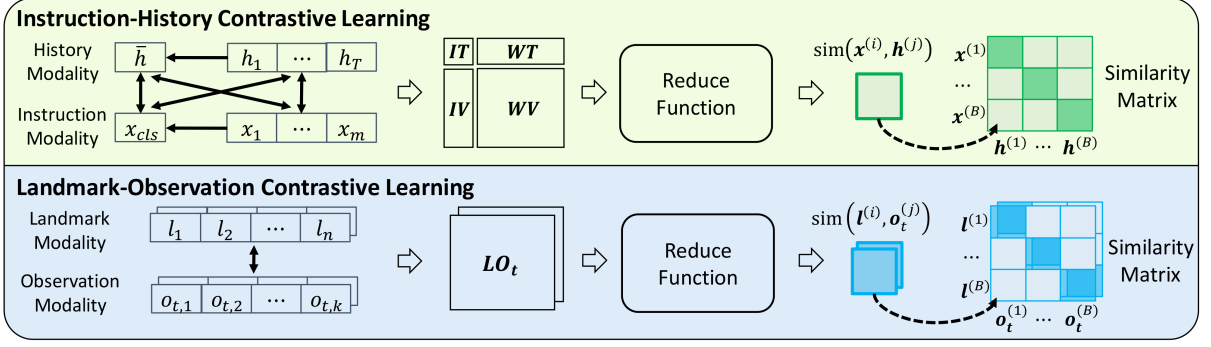


Figure 4: Overview of our DELAN framework. The alignments are employed at both instruction-trajectory and landmark-observation levels. Specifically, we implement instruction-trajectory contrast across global (Instruction, Trajectory) and local (Word, Viewpoint) representations using complete instructions and trajectories. At the landmark-observation level, we perform contrast at each time step.

## 4.2. Instruction-History Alignment

In VLN task, it's crucial to model historical experiences which helps agent make informed decision based on past information. The history modality exhibits a strong semantic correlation with instruction. The instruction comprehensively describes the navigation trajectory while the history records the experience along this trajectory. The instruction token sequence and history token sequence both adhere to chronological order and are synchronized in time. Current unimodal representation of history is solely derived from past observation and action (Chen et al., 2021; Cheng et al., 2022) but lack adequate supervision from textual contents before the fusion stage. Hence, we propose to align the history modality with the instruction part of dual-level instruction. Since they both represent the trajectory and are synchronized in time, exhibiting an element-wise correspondence, we implement the contrastive loss across global representation and local representation. Concretely, we split the instruction-history level alignment into instruction-trajectory alignment, word-trajectory alignment, instruction-viewpoint alignment and word-viewpoint alignment, each providing different benefits.

Additionally, we introduce the global trajectory representation  $\bar{h}$  which obtained through mean pooling of all history embeddings during navigation and choose the textual [CLS] embedding  $x_{cls}$  as the global instruction representation.

**Instruction-Trajectory Alignment** helps global alignment. Given global instruction representation  $x_{cls} \in \mathbb{R}^d$  and global trajectory representation  $\bar{h} \in \mathbb{R}^d$ , the similarity between instruction and trajectory can be computed by matrix multiplication as:

$$S_{IT} = \bar{h}^\top x_{cls}, \quad (7)$$

where  $S_{IT} \in \mathbb{R}^1$  represents for the instruction-trajectory similarity score.

**Word-Trajectory Alignment** assists in emphasizing the most important words and downplaying words with low navigation information, like prepositions and conjunctions. For given local word representation  $X \in \mathbb{R}^{m \times d}$  and global trajectory representation  $\bar{h} \in \mathbb{R}^d$ , we also use matrix multiplication to get the similarity between word and trajectory:

$$S_{WT} = (X\bar{h})^\top, \quad (8)$$

where  $S_{WT} \in \mathbb{R}^{1 \times m}$  represents for the word-trajectory similarity score vector.

**Instruction-Viewpoint Alignment** facilitates generalizing viewpoints which closely follow the instruction. We can compute the similarity between instruction and viewpoint by global instruction representation  $x_{cls} \in \mathbb{R}^d$  and local viewpoint representation (equal to history representation)  $H \in \mathbb{R}^{T \times d}$ , which can be formulated as follows:

$$S_{IV} = Hx_{cls}, \quad (9)$$

where  $S_{IV} \in \mathbb{R}^{T \times 1}$  represents for the instruction-viewpoint similarity score vector.

**Word-Viewpoint Alignment** contributes to monitoring the navigation progress, as each viewpoint embedding can implicitly align with its corresponding sub-instruction word embeddings. The similarity between word and viewpoint can also be obtained from local word representation  $X \in \mathbb{R}^{m \times d}$  and local viewpoint representation  $H \in \mathbb{R}^{T \times d}$  as:

$$S_{WV} = HX^\top, \quad (10)$$

where  $S_{WV} \in \mathbb{R}^{T \times m}$  represents for word-viewpoint similarity score matrix. We employ the reduce function  $R$  in equation 4 to derive scalar values from similarity score vectors or matrices, based on attention mechanisms. Given  $B$  instruction-history pairs  $\{\mathbf{x}^{(i)}, \mathbf{h}^{(i)}\}_{i=1}^B$  where  $\mathbf{x}^{(i)} = \{x_{cls}^{(i)}, X^{(i)}\}$

| Model                        | Val-Seen |             |              |              | Val-Unseen |             |              |              | Test-Unseen |             |              |              |
|------------------------------|----------|-------------|--------------|--------------|------------|-------------|--------------|--------------|-------------|-------------|--------------|--------------|
|                              | TL       | NE↓         | SR↑          | SPL↑         | TL         | NE↓         | SR↑          | SPL↑         | TL          | NE↓         | SR↑          | SPL↑         |
| SF (2018)                    | -        | 3.36        | 66           | -            | -          | 6.62        | 35           | -            | 14.82       | 6.62        | 35           | 28           |
| EnvDrop (2020)               | 11.00    | 3.99        | 62           | 59           | 10.70      | 5.22        | 52           | 48           | 11.66       | 5.23        | 51           | 47           |
| OAAM (2020)                  | 10.20    | -           | 65           | 62           | 9.95       | -           | 54           | 50           | 10.40       | -           | 53           | 50           |
| PREVALENT (2020)             | 10.32    | 3.67        | 69           | 65           | 10.19      | 4.71        | 58           | 53           | 10.51       | 5.30        | 54           | 51           |
| RelGraph (2020a)             | 10.13    | 3.47        | 67           | 65           | 9.99       | 4.73        | 57           | 53           | 10.29       | 4.75        | 55           | 52           |
| SSM (2021)                   | 14.70    | 3.10        | 71           | 62           | 20.70      | 4.32        | 62           | 45           | 20.40       | 4.57        | 61           | 46           |
| VLN <sup>o</sup> BERT (2021) | 11.13    | 2.90        | 72           | 68           | 12.01      | 3.93        | 63           | 57           | 12.35       | 4.09        | 63           | 57           |
| LOViS (2022)                 | -        | 2.40        | 77           | 72           | -          | 3.71        | 65           | 59           | -           | 4.07        | 63           | 58           |
| DUET (2022)                  | 12.32    | 2.28        | 79           | 73           | 13.94      | 3.31        | 72           | 60           | 14.74       | 3.65        | 69           | 59           |
| HAMT (2021)                  | 11.15    | 2.51        | 75.61        | 72.18        | 11.46      | 3.62        | 66.24        | 61.51        | 12.27       | 3.93        | 65.09        | 60.02        |
| HAMT+DELAN                   | 11.16    | 2.42        | 76.89        | 73.61        | 11.62      | 3.46        | 67.22        | 61.53        | 12.34       | 3.96        | 65.30        | 60.70        |
| DUET* (2022)                 | 14.32    | 2.38        | 76.69        | 69.74        | 14.52      | 2.95        | 72.80        | 61.70        | 13.99       | 3.27        | 71.07        | 60.99        |
| DUET+DELAN                   | 12.02    | <b>1.94</b> | <b>81.19</b> | <b>76.66</b> | 13.07      | <b>2.94</b> | <b>73.35</b> | <b>63.39</b> | 12.98       | <b>3.27</b> | <b>71.22</b> | <b>62.69</b> |

Table 1: Comparison of performance with SOTA methods on R2R dataset.

represents in-batch  $i$ -th instruction instance and  $\mathbf{h}^{(i)} = \{\bar{h}^{(i)}, H^{(i)}\}$  represents in-batch  $i$ -th history instance, the score  $\text{sim}(\mathbf{x}^{(i)}, \mathbf{h}^{(j)})$  measures the semantic similarity between two instances and can be averaged from various similarity scores:

$$\text{sim}(\mathbf{x}^{(i)}, \mathbf{h}^{(j)}) = (S_{IT} + R(S_{WT}) + R(S_{IV}) + R(S_{WV}))/4 \quad (11)$$

Finally, the instruction-history level contrastive loss can be calculated from equation 5 and equation 11:

$$\mathcal{L}_{IH} = L(\mathbf{x}, \mathbf{h}) \quad (12)$$

### 4.3. Landmark-Observation Alignment

The panoramic observation at each time step consists of  $k$  view images together with relative directions, providing a detailed spatial layout. Such observation exhibits strong semantic correlation with landmarks in instructions, as the navigation can be typically succeeded by following landmark entities in the environment based on landmark references in instructions. The pre-fusion alignment between landmark and observation enhance the interaction between related representations.

Consequently, we align the observation modality with the landmark part of dual-level instructions. The observation and landmarks exhibit an element-wise correspondence, as landmarks only present in some of the view images and real-time observation contains only a subset of the landmarks mentioned in instruction. Therefore, we implement the contrastive loss only across local representation here.

**Landmark-Observation Alignment** helps align landmark tokens with specific view images of observation at each time step. The similarity matrix can be obtained from local landmark representation  $L \in \mathbb{R}^{n \times d}$  and local representation observation  $O_t \in \mathbb{R}^{k \times d}$  at time step  $t$  as follows:

$$S_{LO_t} = O_t L^T, \quad (13)$$

where  $S_{LO_t} \in \mathbb{R}^{k \times n}$  is the landmark-observation similarity score matrix at step  $t$ . We also employ the reduce function  $R$  in equation 4 to dynamically score the importance of each element in similarity matrix and aggregate them. Analogously, given  $B$  landmark-observation pairs  $\{\mathbf{l}^{(i)}, \mathbf{o}_t^{(i)}\}_{i=1}^B$  where  $\mathbf{l}^{(i)} = \{L^{(i)}\}$  represents in-batch  $i$ -th landmark instance and  $\mathbf{o}_t^{(i)} = \{O_t^{(i)}\}$  represents in-batch  $i$ -th observation instance at time step  $t$ , the similarity score  $\text{sim}(\mathbf{l}^{(i)}, \mathbf{o}_t^{(j)})$  measures the semantic similarity between the landmark and observation:

$$\text{sim}(\mathbf{l}^{(i)}, \mathbf{o}_t^{(j)}) = R(S_{LO_t}) \quad (14)$$

Finally, the landmark-observation level contrastive loss can be computed from averaged similarity scores along the trajectory according to equation 5 and equation 14. The formula is:

$$\mathcal{L}_{LO} = \frac{1}{T} \sum_{t=1}^T L(\mathbf{l}, \mathbf{o}_t) \quad (15)$$

Note that  $T$  is the trajectory length.

### 4.4. Training Strategy

Two typical training strategies in VLN task are Imitation Learning (IL) and Reinforcement Learning (RL). In IL phase, the agent learns by following teacher action  $a_t^*$ :  $\mathcal{L}_{IL} = -\frac{1}{T} \sum_{t=1}^T a_t^* \log(p_t)$ , where  $p_t$  is predicted action probability. In RL phase, the agent samples the action to explore the environment and learns from the reward:  $\mathcal{L}_{RL} = -\frac{1}{T} \sum_{t=1}^T a_t^s \log(p_t) A_t$ , where  $a_t^s$  is the sampled action according to predicted action probability  $p_t$  and  $A_t$  is the advantage in A2C algorithm (Mnih et al., 2016). Overall, by combining contrastive loss on instruction-history level and landmark-observation level, the total loss is calculated as:

$$\mathcal{L}_{loss} = \lambda_1 \mathcal{L}_{RL} + \lambda_2 \mathcal{L}_{IL} + \lambda_3 \mathcal{L}_{IH} + \lambda_4 \mathcal{L}_{LO} \quad (16)$$

| Model                        | Val-Seen    |             |             |             | Val-Unseen  |             |             |             | Test-Unseen |             |             |             |
|------------------------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|
|                              | SR↑         | SPL↑        | nDTW↑       | sDTW↑       | SR↑         | SPL↑        | nDTW↑       | sDTW↑       | SR↑         | SPL↑        | nDTW↑       | sDTW↑       |
| Multilingual Baseline (2020) | 25.2        | -           | 42.2        | 20.7        | 22.8        | -           | 38.9        | 18.2        | 21.0        | 18.6        | 41.1        | 20.6        |
| Monolingual Baseline (2020)  | 28.8        | -           | 46.8        | 23.8        | 28.5        | -           | 44.5        | 23.1        | 25.4        | 22.6        | 41.1        | 32.4        |
| EnvDrop-CLIP (2020)          | -           | -           | -           | -           | 42.6        | -           | 55.7        | -           | 38.3        | -           | 51.1        | 32.4        |
| Multilingual HAMT (2021)     | 59.4        | <b>58.9</b> | 65.3        | 50.9        | 56.5        | <b>56.0</b> | 63.1        | 48.3        | 53.1        | 46.6        | 59.9        | 45.2        |
| Multilingual HAMT+DELAN      | <b>62.1</b> | 58.5        | <b>67.9</b> | <b>53.4</b> | <b>57.5</b> | 53.9        | <b>65.4</b> | <b>49.5</b> | <b>53.9</b> | <b>47.7</b> | <b>60.9</b> | <b>45.9</b> |

Table 2: Comparison of performance with SOTA methods on RxR dataset.

| Model            | Val-Unseen  |             |             |             |             |
|------------------|-------------|-------------|-------------|-------------|-------------|
|                  | NE↓         | SR↑         | CLS↑        | nDTW↑       | sDTW↑       |
| SF (2018)        | 8.47        | 24.0        | 30.0        | -           | -           |
| RCM (2019)       | 8.08        | 26.0        | 35          | 30          | 13          |
| OAAM (2020)      | 8.51        | 26.6        | 36.2        | 30.3        | 12.6        |
| RelGraph (2020a) | 7.43        | 36.0        | 41.0        | 47.0        | 34.0        |
| VLN+BERT (2021)  | 6.48        | 43.6        | 51.4        | 45.1        | 29.9        |
| DDL (2022)       | 6.43        | 42.4        | 43.6        | 38.5        | 21.0        |
| LOViS (2022)     | 6.07        | 45.0        | 45.0        | 43.0        | 23.0        |
| DUET (2022)      | 5.60        | 51.2        | 47.0        | 42.7        | 29.3        |
| HAMT (2021)      | 6.09        | 44.6        | 57.7        | 50.3        | 31.8        |
| HAMT+DELAN       | 5.99        | 45.5        | <b>61.0</b> | <b>54.0</b> | <b>34.4</b> |
| DUET* (2022)     | <b>5.42</b> | 51.4        | 46.0        | 40.2        | 27.1        |
| DUET+DELAN       | 5.47        | <b>52.5</b> | 47.6        | 41.6        | 28.9        |

Table 3: Comparison of performance with SOTA methods on R4R dataset.

where  $\lambda_1$ ,  $\lambda_2$ ,  $\lambda_3$  and  $\lambda_4$  are weighting parameters that balance different losses.

## 5. Experiments

### 5.1. Experimental Setup

**Datasets.** We evaluated our method on three representative VLN tasks (four datasets): VLN with fine-grained instructions R2R (Anderson et al., 2018b) and RxR (Ku et al., 2020); long-horizon VLN R4R (Jain et al., 2019); and vision-and-dialogue navigation CVDN (Thomason et al., 2020).

- **R2R** is the basic VLN dataset which contains step-by-step navigation instructions. It includes 90 scans with 22k human-annotated instruction-trajectory pairs. The dataset is divided into train, val seen, val unseen and test unseen splits, consisting of 61, 56, 11 and 18 scans respectively. Scans in the unseen splits are excluded from the training set.
- **RxR** is a multilingual (English, Hindi and Telugu) VLN dataset. Comparing to R2R, RxR is 10x larger with longer and more various paths.
- **R4R** extends R2R dataset by concatenating two adjacent trajectories and their instructions. It encourages the agent to navigate as human plan instead of going to the destination directly.
- **CVDN** requires an agent to navigate based on multi-turn question-answering dialogues. The lengths of instructions and trajectories are much longer than R2R.

| Model               | Val-Seen    | Val-Unseen  | Test-Unseen |
|---------------------|-------------|-------------|-------------|
| PREVALENT (2020)    | -           | 3.15        | 2.44        |
| VISITRON (2021)     | 5.11        | 3.25        | 3.11        |
| MT-RCM+EnvAg (2020) | 5.07        | 4.65        | 3.91        |
| MTVM (2022)         | -           | 5.15        | 4.82        |
| HAMT (2021)         | 6.91        | 5.13        | 5.58        |
| HAMT+DELAN          | 7.18        | 5.20        | 5.85        |
| DUET* (2022)        | <b>8.57</b> | 5.58        | 5.96        |
| DUET+DELAN          | 7.40        | <b>5.95</b> | <b>6.19</b> |

Table 4: Comparison of goal progress with SOTA methods on CVDN dataset.

**Evaluation Metrics.** For R2R, we use standard evaluation metrics following previous work (Anderson et al., 2018a,b): Trajectory Length (TL), Navigation Error (NE), Success Rate (SR), Success rate weighted by Path Length (SPL). Metrics that measure the path fidelity between the predicted path and target path are introduced for R4R and RxR, including normalized Dynamic Time Warping (nDTW), Success weighted by normalized Dynamic Time Warping (sDTW) (Ilharco et al., 2019) and Coverage weighted by Length Score (CLS) (Jain et al., 2019). For CVDN, we use Goal Progress (GP) in meters (Thomason et al., 2020) as the primary metric. GP measures the difference between completed distance and left distance to the goal, so the higher the better. Note that **bold** metrics in tables indicate the best results.

**Implementation Details.** The experiments are conducted on a NVIDIA 3090 GPU. For contrastive loss, we set  $\lambda_3 = 0.01$  and  $\lambda_4 = 0.1$  in equation 16 while keeping temperature parameter  $\tau = 1$ . We also introduce the memory bank (Wu et al., 2018) to expand number of negative samples and fix the size to 480. Our baseline models consist of HAMT (Chen et al., 2021) and DUET (Chen et al., 2022). Different from the original implementation (Chen et al., 2022), we utilize CLIP-ViT-B-16 (Li et al., 2022a) as vision encoder for DUET. The batch size is set to 4 for DUET and 8 for HAMT. The optimizer is AdamW (Loshchilov and Hutter, 2017). The remaining training schedules remain consistent with the baselines. Inference is performed using a greedy search approach, following a single-run setting (Wang et al., 2019). Note that models marked with \* in table denote our re-implementation.

| Model | # | Alignment | Val-Unseen   |             |              |              |
|-------|---|-----------|--------------|-------------|--------------|--------------|
|       |   |           | TL           | NE          | SR           | SPL          |
| DUET  | 1 | Dual      | <b>13.07</b> | <b>2.94</b> | <b>73.35</b> | <b>63.39</b> |
|       | 2 | Single    | 14.42        | 2.95        | 72.88        | 61.78        |

Table 5: Ablation study on alignment level on the R2R dataset.

| Model                    | # | Val-Unseen   |             |              |              |
|--------------------------|---|--------------|-------------|--------------|--------------|
|                          |   | TL           | NE          | SR           | SPL          |
| DUET+DELAN               | 1 | <b>13.07</b> | 2.94        | <b>73.35</b> | <b>63.39</b> |
| w/o Instr-Land Attention | 2 | 15.52        | 3.05        | 72.33        | 60.42        |
| w/o Shared Encoder       | 3 | 16.33        | <b>2.92</b> | 72.67        | 59.35        |

Table 6: Ablation study on separation strategies for instructions and landmarks on the R2R dataset.

## 5.2. Main Results

**Fine-grained VLN: R2R and RxR.** Table 1 compares the performances of different agents on R2R benchmark. Our method outperforms baselines over all dataset splits, achieving 62.69% SPL (+1.7%) on the test split. Table 2 shows results compared with state-of-the-art methods on RxR dataset. Our method gets significant improvements (+1.1% on SPL and +1.0% on SR) on test split compared with the backbones. All above demonstrate the effectiveness and generalization of our framework.

**Long-horizon VLN: R4R.** Table 3 shows navigation results on R4R dataset. DELAN performs consistently better than the baseline HAMT (+0.9% on SR), especially on these path fidelity related metrics (+3.3% on CLS, +3.7% on nDTW and +2.6% on sDTW). This demonstrates its strong instruction following ability in long-horizon navigation scenario. DELAN can model relation between instruction and history to monitor progress, and closely follow the abundant landmarks in long instruction to navigate. The graph-based baseline DUET naturally generates more sub-trajectories for back-and-forth exploration in the predicted path, resulting in reduced path fidelity metrics (CLS, nDTW and sDTW). However, our method can still improve almost all of its navigation metrics (especially +1.1% on SR), which explains the universality of our framework.

**Vision-and-Dialog Navigation: CVDN.** The results in Table 4 demonstrate that our method enhances the goal progress of previous models, increasing HAMT’s performance by 0.27 meters and DUET’s by 0.23 meters in test environments. This highlights the effectiveness of our proposed framework in a dialog setting.

| Model | # | Component |    |         |    | Val-Unseen |       |             |              |              |
|-------|---|-----------|----|---------|----|------------|-------|-------------|--------------|--------------|
|       |   | Instr-His |    | Land-Ob |    | TL         | NE↓   | SR↑         | SPL↑         |              |
|       |   | IT        | WT | IV      | WV |            |       |             |              | LO           |
| DUET  | 1 | ✓         | ✓  | ✓       | ✓  | ✓          | 13.07 | 2.94        | <b>73.35</b> | <b>63.39</b> |
|       | 2 |           |    |         |    |            | 12.99 | 3.02        | 71.95        | 61.97        |
|       | 3 | ✓         | ✓  | ✓       | ✓  | ✓          | 14.58 | 2.96        | 72.84        | 60.65        |
|       | 4 | ✓         | ✓  | ✓       | ✓  | ✓          | 12.84 | 3.11        | 71.56        | 62.44        |
|       | 5 | ✓         | ✓  | ✓       | ✓  | ✓          | 13.31 | 2.94        | 73.09        | 62.96        |
|       | 6 | ✓         | ✓  | ✓       | ✓  |            | 14.25 | <b>2.93</b> | 72.88        | 61.19        |
|       | 7 |           |    |         |    | ✓          | 13.06 | 2.94        | 72.84        | 63.01        |

Table 7: Ablation study on different components of dual-level alignment on the R2R dataset.

## 5.3. Ablation Study

**Dual-level Alignment.** We first evaluate the necessity of employing alignment with dual levels rather than single level, as shown in Table 5. Line 1 represents our DELAN framework with dual-level alignment and corresponding dual-level instruction. Line 2 depicts the use of single-level alignment, which aligns only the original instruction with concatenated visual-related components (history and observation). The results indicate that explicit alignment at dual levels significantly enhance navigation performance compared to directly aligning vision and language modalities.

**Separation of Instruction and Landmark.** In the DELAN framework, we adopt a method of separating landmark words from instructions to facilitate explicit modality alignment. Table 6 presents performance under three approaches with increasing degrees of separation between instructions and landmarks. Line 1 illustrates the strategy employed in DELAN, allowing instructions and landmarks attending to each other using the shared text encoder. Line 2 represents a strategy where instructions and landmarks are encoded separately but with the shared text encoder. Line 3 represents a strategy utilizing two independent encoders to encode instructions and landmarks respectively, ensuring complete instruction and landmark separation before the fusion stage. The results presented in Table 6 highlight the superiority of the separation approach in DELAN. Since individual landmark words still hold contextual relevance, allowing them to attend to instruction words makes practical sense.

**Components of DELAN.** We study the effectiveness of each component of DELAN over the R2R val unseen split. As shown in Table 7, both instruction-history (Instr-His) and landmark-observation (Land-Ob) alignment contributes to the excellent performance of DELAN, especially on metric NE and SR. Line 2 to 5 in Table 7 assess the influence of four components of the instruction-history level contrastive learning, i.e. instruction-trajectory (IT), word-trajectory (WT), instruction-viewpoint (IV), and word-viewpoint (WV). Remov-



| Model          | Stop         | Turn         | Object       | Room         | Avg.         |
|----------------|--------------|--------------|--------------|--------------|--------------|
| HAMT (2021)    | 70.62        | 39.61        | 12.01        | 24.96        | 36.80        |
| HAMT+Instr-His | 69.38        | 38.02        | 13.16        | 26.72        | 36.82        |
| HAMT+Land-Ob   | 65.73        | 38.60        | 14.74        | <b>27.89</b> | 36.74        |
| HAMT+DELAN     | <b>71.27</b> | <b>40.73</b> | <b>15.61</b> | 25.59        | <b>38.30</b> |

Table 8: Quantitative analysis on the mastery of navigation skills on RxR dataset. Instr-His indicates the instruction-history alignment and Land-Ob indicates the landmark-observation alignment.

ing the word-trajectory alignment significantly increases the trajectory length, indicating a decrease in the overall understanding of navigation progress. The elimination of instruction-viewpoint alignment degrade the success rate most, showing the importance of aligning history viewpoints with instructions. Each component has a positive effect on the performance. These results demonstrate that DELAN is not only effective but also reasonable.

#### 5.4. Further Analysis

We further analyze the agent’s behavior in four specific navigation skills described in Yang et al. (2023). These skills are **Stop**, **Turn**, **Object** and **Room**, indicating the agent’s ability in stopping, changing direction, object seeking, and room finding, respectively. A higher score means a stronger ability. As shown in Table 8, our DELAN framework significantly enhances the performance across all skills (average +1.5). We attribute the notable promotion in object seeking (+3.6) to our proposed landmark-observation level alignment, which enforces the agent to recognize landmarks appeared in panoramic observations. Meanwhile, the ability of agents to stop and change directions has also been improved, indicating that under the DELAN framework, the agent has a better judgment of instruction execution and navigation progress.

#### 5.5. Qualitative Results

We visualize the trajectories of DUET (Chen et al., 2022) and DUET+DELAN in Figure 5. Due to the lack of the pre-fusion alignment, DUET has the unsatisfactory instruction-following capacity. For example, it fails to follow the instruction "turn slight right" or misses important landmark "painting placed on the wall", leading to either incorrect endpoints (upper) or unnecessary exploration trajectories (lower). In contrast, our proposed DELAN framework could promote the instruction comprehension and help the agent make correct decisions, resulting in a more precise and efficient navigation.

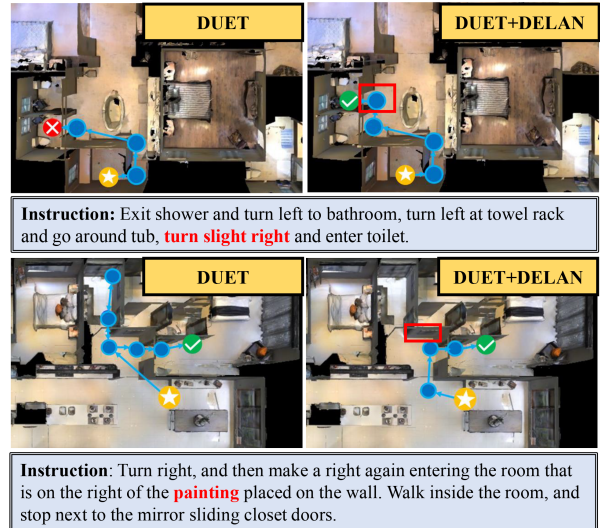


Figure 5: Trajectories of DUET and DUET+DELAN in R2R-unseen environments. Yellow, green, and red nodes signify the start, target, and incorrect endpoints respectively. Red rectangles correspond the red pats of the instructions.

## 6. Conclusion

This paper highlights the significance of pre-fusion alignment in the VLN task and introduces a dual-level alignment (DELAN) framework, focusing on the instruction-history and landmark-observation levels. For adaptation to the dual-level alignment, we reconstruct a dual-level instruction through concatenating original instruction and landmarks. We employ different self-supervised contrastive learning strategies to realize the dual-level alignment respectively. Experiments across various VLN benchmarks demonstrate the effectiveness, generalization and universality of our approach. We hope this work can inspire further explorations on modality alignment in VLN and related fields.

## 7. Acknowledgement

This work is supported by National Natural Science Foundation of China (No.71991471, No.6217020551) and Science and Technology Commission of Shanghai Municipality Grant (No.21DZ1201402).

## 8. Bibliographical References

Peter Anderson, Angel Chang, Devendra Singh Chaplot, Alexey Dosovitskiy, Saurabh Gupta, Vladlen Koltun, Jana Kosecka, Jitendra Malik, Roozbeh Mottaghi, Manolis Savva, et al. 2018a.

- On evaluation of embodied navigation agents. *arXiv preprint arXiv:1807.06757*.
- Peter Anderson, Qi Wu, Damien Teney, Jake Bruce, Mark Johnson, Niko Sünderhauf, Ian Reid, Stephen Gould, and Anton Van Den Hengel. 2018b. 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.
- Shizhe Chen, Pierre-Louis Guhur, Cordelia Schmid, and Ivan Laptev. 2021. History aware multimodal transformer for vision-and-language navigation. *Advances in neural information processing systems*, 34:5834–5847.
- Shizhe Chen, Pierre-Louis Guhur, Makarand Tapaswi, Cordelia Schmid, and Ivan Laptev. 2022. 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 (CVPR)*, pages 16537–16547.
- Wenhao Cheng, Xingping Dong, Salman Khan, and Jianbing Shen. 2022. Learning disentanglement with decoupled labels for vision-language navigation. In *Computer Vision—ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXVI*, pages 309–329. Springer.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.
- Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2021. [An image is worth 16x16 words: Transformers for image recognition at scale](#). In *International Conference on Learning Representations*.
- Chenzhuang Du, Jiaye Teng, Tingle Li, Yichen Liu, Tianyuan Yuan, Yue Wang, Yang Yuan, and Hang Zhao. 2023. On uni-modal feature learning in supervised multi-modal learning. *arXiv preprint arXiv:2305.01233*.
- Daniel Fried, Ronghang Hu, Volkan Cirik, Anna Rohrbach, Jacob Andreas, Louis-Philippe Morency, Taylor Berg-Kirkpatrick, Kate Saenko, Dan Klein, and Trevor Darrell. 2018. Speaker-follower models for vision-and-language navigation. *Advances in Neural Information Processing Systems*, 31.
- Jing Gu, Eliana Stefani, Qi Wu, Jesse Thomason, and Xin Eric Wang. 2022. Vision-and-language navigation: A survey of tasks, methods, and future directions. *arXiv preprint arXiv:2203.12667*.
- Weituo Hao, Chunyuan Li, Xiujun Li, Lawrence Carin, and Jianfeng Gao. 2020. Towards learning a generic agent for vision-and-language navigation via pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13137–13146.
- Yicong Hong, Cristian Rodriguez, Yuankai Qi, Qi Wu, and Stephen Gould. 2020a. Language and visual entity relationship graph for agent navigation. *Advances in Neural Information Processing Systems*, 33:7685–7696.
- Yicong Hong, Cristian Rodriguez-Opazo, Qi Wu, and Stephen Gould. 2020b. Sub-instruction aware vision-and-language navigation. *arXiv preprint arXiv:2004.02707*.
- Yicong Hong, Qi Wu, Yuankai Qi, Cristian Rodriguez-Opazo, and Stephen Gould. 2021. 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.
- Gabriel Ilharco, Vihan Jain, Alexander Ku, Eugene Ie, and Jason Baldridge. 2019. General evaluation for instruction conditioned navigation using dynamic time warping. *arXiv preprint arXiv:1907.05446*.
- Vihan Jain, Gabriel Magalhaes, Alexander Ku, Ashish Vaswani, Eugene Ie, and Jason Baldridge. 2019. Stay on the path: Instruction fidelity in vision-and-language navigation. *arXiv preprint arXiv:1905.12255*.
- Alexander Ku, Peter Anderson, Roma Patel, Eugene Ie, and Jason Baldridge. 2020. Room-across-room: Multilingual vision-and-language navigation with dense spatiotemporal grounding. *arXiv preprint arXiv:2010.07954*.
- Jialu Li, Hao Tan, and Mohit Bansal. 2022a. Envedit: Environment editing for vision-and-language navigation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15407–15417.
- Junnan Li, Ramprasaath Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and Steven Chu Hong Hoi. 2021. Align before fuse: Vision and language representation learning with momentum distillation. *Advances in neural information processing systems*, 34:9694–9705.

- Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, et al. 2020. Oscar: Object-semantics aligned pre-training for vision-language tasks. In *European Conference on Computer Vision*, pages 121–137. Springer.
- Zejun Li, Zhihao Fan, Huaixiao Tou, Jingjing Chen, Zhongyu Wei, and Xuanjing Huang. 2022b. Mvptr: Multi-level semantic alignment for vision-language pre-training via multi-stage learning. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 4395–4405.
- Xiwen Liang, Fengda Zhu, Yi Zhu, Bingqian Lin, Bing Wang, and Xiaodan Liang. 2022. Contrastive instruction-trajectory learning for vision-language navigation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 1592–1600.
- Chuang Lin, Yi Jiang, Jianfei Cai, Lizhen Qu, Gholamreza Haffari, and Zehuan Yuan. 2022. Multi-modal transformer with variable-length memory for vision-and-language navigation. In *Computer Vision—ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXVI*, pages 380–397. Springer.
- Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.
- Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *Advances in neural information processing systems*, 32.
- Chih-Yao Ma, Jiasen Lu, Zuxuan Wu, Ghassan AlRegib, Zsolt Kira, Richard Socher, and Caiming Xiong. 2019a. Self-monitoring navigation agent via auxiliary progress estimation. *arXiv preprint arXiv:1901.03035*.
- Chih-Yao Ma, Zuxuan Wu, Ghassan AlRegib, Caiming Xiong, and Zsolt Kira. 2019b. The regretful agent: Heuristic-aided navigation through progress estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6732–6740.
- Yiwei Ma, Guohai Xu, Xiaoshuai Sun, Ming Yan, Ji Zhang, and Rongrong Ji. 2022. X-clip: End-to-end multi-grained contrastive learning for video-text retrieval. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 638–647.
- Volodymyr Mnih, Adria Puigdomenech Badia, Mehdi Mirza, Alex Graves, Timothy Lillicrap, Tim Harley, David Silver, and Koray Kavukcuoglu. 2016. Asynchronous methods for deep reinforcement learning. In *International conference on machine learning*, pages 1928–1937. PMLR.
- Yuankai Qi, Zizheng Pan, Shengping Zhang, Anton van den Hengel, and Qi Wu. 2020. Object-and-action aware model for visual language navigation. In *European Conference on Computer Vision*, pages 303–317. Springer.
- Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR.
- Ayush Shrivastava, Karthik Gopalakrishnan, Yang Liu, Robinson Piramuthu, Gokhan Tür, Devi Parikh, and Dilek Hakkani-Tür. 2021. Visitron: Visual semantics-aligned interactively trained object-navigator. *arXiv preprint arXiv:2105.11589*.
- Hao Tan, Licheng Yu, and Mohit Bansal. 2019. Learning to navigate unseen environments: Back translation with environmental dropout. *arXiv preprint arXiv:1904.04195*.
- Hao Hao Tan and Mohit Bansal. 2019. Lxmert: Learning cross-modality encoder representations from transformers. *ArXiv*, abs/1908.07490.
- Jesse Thomason, Michael Murray, Maya Cakmak, and Luke Zettlemoyer. 2020. Vision-and-dialog navigation. In *Conference on Robot Learning*, pages 394–406. PMLR.
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017a. Attention is all you need. *Advances in neural information processing systems*, 30.
- Ashish Vaswani, Noam M. Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017b. Attention is all you need. *ArXiv*, abs/1706.03762.
- Hanqing Wang, Wenguan Wang, Wei Liang, Caiming Xiong, and Jianbing Shen. 2021. Structured scene memory for vision-language navigation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8455–8464.
- Xin Wang, Qiuyuan Huang, Asli Celikyilmaz, Jianfeng Gao, Dinghan Shen, Yuan-Fang Wang, William Yang Wang, and Lei Zhang. 2019. Reinforced cross-modal matching and self-supervised imitation learning for vision-language

- navigation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6629–6638.
- Xin Eric Wang, Vihan Jain, Eugene Ie, William Yang Wang, Zornitsa Kozareva, and Sujith Ravi. 2020. Environment-agnostic multitask learning for natural language grounded navigation. In *Computer Vision—ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXIV 16*, pages 413–430. Springer.
- Zhirong Wu, Yuanjun Xiong, Stella X Yu, and Dahua Lin. 2018. Unsupervised feature learning via non-parametric instance discrimination. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3733–3742.
- Jianwei Yang, Yonatan Bisk, and Jianfeng Gao. 2021. Taco: Token-aware cascade contrastive learning for video-text alignment. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11562–11572.
- Zijiao Yang, Arjun Majumdar, and Stefan Lee. 2023. Behavioral analysis of vision-and-language navigation agents. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2574–2582.
- Lewei Yao, Runhui Huang, Lu Hou, Guansong Lu, Minzhe Niu, Hang Xu, Xiaodan Liang, Zhenguo Li, Xin Jiang, and Chunjing Xu. 2021. Filip: fine-grained interactive language-image pre-training. *arXiv preprint arXiv:2111.07783*.
- Jiwen Zhang, Zhongyu Wei, Jianqing Fan, and Ji-ajie Peng. 2021. Curriculum learning for vision-and-language navigation. In *NeurIPS*.
- Yue Zhang and Parisa Kordjamshidi. 2022. Lo-vis: Learning orientation and visual signals for vision and language navigation. *arXiv preprint arXiv:2209.12723*.
- Fengda Zhu, Yi Zhu, Xiaojun Chang, and Xiaodan Liang. 2020. Vision-language navigation with self-supervised auxiliary reasoning tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10012–10022.