

PRET: Planning with Directed Fidelity Trajectory for Vision and Language Navigation

Renjie Lu¹, Jingke Meng¹, and Wei-Shi Zheng^{1,2,3}

¹ School of Computer Science and Engineering, Sun Yat-sen University, Guangzhou, China

lurj3@mail2.sysu.edu.cn, mengjke@gmail.com, wszheng@ieee.org

² Peng Cheng Laboratory, Shenzhen, China

³ Key Laboratory of Machine Intelligence and Advanced Computing, Ministry of Education, Guangzhou, China

Abstract. Vision and language navigation is a task that requires an agent to navigate according to a natural language instruction. Recent methods predict sub-goals on constructed topology map at each step to enable long-term action planning. However, they suffer from high computational cost when attempting to support such high-level predictions with GCN-like models. In this work, we propose an alternative method that facilitates navigation planning by considering the alignment between instructions and directed fidelity trajectories, which refers to a path from the initial node to the candidate locations on a directed graph without detours. This planning strategy leads to an efficient model while achieving strong performance. Specifically, we introduce a directed graph to illustrate the explored area of the environment, emphasizing directionality. Then, we firstly define the trajectory representation as a sequence of directed edge features, which are extracted from the panorama based on the corresponding orientation. Ultimately, we assess and compare the alignment between instruction and different trajectories during navigation to determine the next navigation target. Our method outperforms previous SOTA method BEVBert on RxR dataset and is comparable on R2R dataset while largely reducing the computational cost. Code is available: <https://github.com/iSEE-Laboratory/VLN-PRET>.

Keywords: Vision-and-Language Navigation · Planning

1 Introduction

Enabling a robot to perform tasks on behalf of humans has been a longstanding objective in AI research. One such task is vision-and-language navigation (VLN) [4, 23, 36, 41], where an agent is required to navigate to a desired location by following natural language instructions provided by humans. For example, given the instruction “Go up stairs and stop at the top in front of a mirror.”, the agent needs to follow the instruction and stop at an appropriate location. VLN has attracted numerous research interests [2, 7, 8, 14, 16–18, 20, 26, 31, 40, 44, 47].

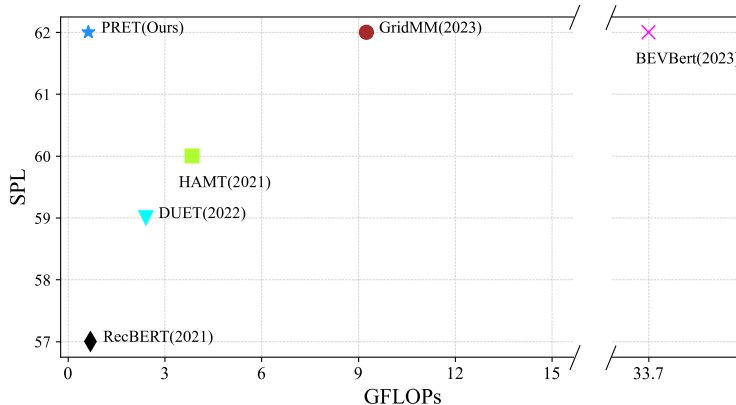


Fig. 1: Comparison of SPL [3] and GFLOPs on R2R test unseen split dataset. Our method is comparable with previous SOTA methods while being more computational efficient. The computational cost of text encoder and visual encoder is omitted for fair comparison.

Recent methods [2, 8, 16, 20, 44] demonstrate the effectiveness of introducing maps to enlarge the decision space to improve planning strategies. Instead of predicting low-level actions limited to short movements, these methods construct a graph to keep track of all visited and navigable locations observed so far, which enables high-level planning by expanding the action space to encompass the entire explored area. With the constructed topological maps, these methods predict actions in the global space via GCN-like models, where node features aggregate neighboring information. However, repeatedly calculating the entire map to predict actions at every step, even if the topo-map has only minor changes, is inefficient. In addition, formulating vision information of an environment in graph structure is too coarse-grained for accurate decision-making. Previous methods [2,8] address this problem by introducing an additional branch to incorporate fine-grained information, such as local image features and bird’s-eye-view features. But this strategy increases the model complexity and further escalates computational cost.

In this work, we present an alternative way that supports the global decision space while achieving comparable performance. Our method, named **P**lanning with **D**i**R**ected **F**id**E**lity **T**rajectory(PRET), does not require calculating the entire graph at each step, nor does it rely on incorporating additional fine-grained information. The main idea is to determine the next location to navigate by evaluating the alignment between the instruction and the visual observations along different trajectories between the start point to all the unvisited nodes, as shown in Figure 2(b). Specifically, we maintain a directed fidelity trajectory(colored in red) for each unvisited node. The fidelity trajectory refers to a path from initial node to an unvisited node without detours. We assess the alignment between each trajectory and the instruction, and select the unvisited node with the highest alignment score to navigate. This planning strategy is efficient as we only need to compute for newly observed nodes. We estimate the instruction-trajectory

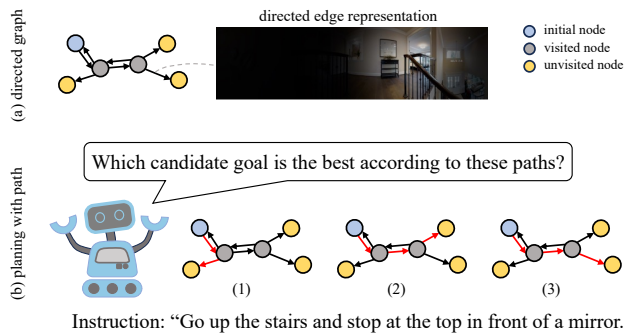


Fig. 2: Illustration of our approaches. (a) shows our directed graph representation. Each edge is assigned with an orientation-aware panorama feature. (b) depict our planning method. We select an unvisited node (colored yellow) to navigate towards next by choosing the fidelity path (colored red) that best aligned with instruction.

alignment with transformer and further reduce the computational cost by utilizing the KV-cache [35] technique. It also enables us to make decisions by taking into full consideration of the instruction-trajectory alignment.

In order to enhance the alignment between instructions and navigation trajectories, we propose to incorporate directionality in path representation. Due to the inherent directional nature of the navigation process, it is crucial that the representation of trajectories in opposite directions is asymmetric, as visual observations are linked to observed orientation. In this work, we introduce the directed graph to depict the explored area of the environment and **firstly** define the trajectory representation as a sequence of directed edge features. These directed edge features are derived from the panorama based on the corresponding orientation (as depicted in Fig. 2(a)). By incorporating directionality in the graph edge features, we eliminate irrelevant panorama information and obtain a more accurate representation of the directed trajectory. Our method overcome existing methods [2, 8, 16] that commonly adopt the undirected graph to store panorama features on graph nodes, resulting in direction-invariant path representations, which fails to capture the important distinction between paths that traverse the same node from different directions. For instance, consider a scenario where agents pass through the same node from different directions. Despite the fundamentally different spatial contexts of these paths, the node representation remains unchanged.

We conduct experiments on R2R, RxR datasets to evaluate the efficacy of our proposed methods. The results demonstrate that PRET achieves strong performance while being more efficient than previous methods. Specifically, PRET achieves comparable performance compared to previous state-of-the-art method BEVBert [2] on R2R dataset with only 3% computational cost as shown in Fig. 1. On the RxR dataset, PRET outperforms previous methods and achieves the new state-of-the-art performance. Qualitative visualization also shows that our simplified model is able to learn complex backtracking strategies.

2 Related Work

Vision-and-Language Navigation(VLN). Early VLN methods use sequence-to-sequence LSTMs with attention mechanism to encode language features and predict local actions. SpeakerFollower [14] first introduces back translation [38] technique to ease the data scarcity problem in VLN. EnvDrop [40] proposes to augment environment by using dropout on feature space to avoid overfitting on visual input. Following this work, many environment augmentation approaches [24, 29] is proposed. Some auxiliary tasks is proposed by [31, 48] for better guidance. More recently, transformer is explored how to be adopted in VLN. PRESS [27] uses BERT [11] as the text encoder. PREVALENT [17] first propose to pretrain the transformer in VLN dataset, but still use it as encoder. More recent works [7, 18, 28] explore how to memorize navigation history and use transformer to learn strong planning strategy.

Navigation Strategy in VLN. Navigation strategy plays a key role in Vision-and-Language Navigation, as VLN requires agents to navigate in unseen environments and thus needs to explore and familiarize the environment. Works like [32] investigate designing regretful agents to enable explicit exploration. Reinforcement learning methods [40, 46] have also been explored to enhance navigation strategies. SSM [44] constructs directed graphs to represent explored areas, with vision features on nodes and orientation on edges. DUET [8] employs a dual-scale transformer to make local and global predictions. AZHP [16] constructs hierarchy graphs to facilitate exploration. MetaExplore [20] explicitly predicts whether to backtrack using a separate module.

Maps For Navigation. Navigation research has a long history of using SLAM [15] to construct maps [5, 6, 42] for planning. In the VLN literature, several methods [8, 44] adopt topological graphs for global planning. However, the visual representation on topological maps is too coarse-grained for decision-making. To address this limitation, [2, 47] introduce metric maps in VLN. While grid-based metric maps can precisely represent spatial layouts, they result in high computational costs. To balance representation ability and computational cost, BEVBert [2] utilizes learnable hybrid topological and metric maps. However, the computational cost remains high. In our work, we propose a novel approach that involves constructing a directed topo-map and planning with trajectory. Our method is more efficient as we perform planning with trajectory instead of directly relying on GCN-like models to encode the entire map for planning. This also allows us to incrementally calculate embeddings for new trajectories and further reduces computational cost.

3 Method

3.1 Problem Formulation

In VLN with discrete environments [4, 23, 36, 41], an environment is an undirected graph $\mathcal{G} = \{\mathcal{V}, \mathcal{E}\}$, where $\mathcal{V} = \{V_i\}_{i=1}^N$ represents N navigable nodes and \mathcal{E} denotes navigable edges. At the beginning of navigation, an agent is initialized

at a starting node and given a natural language instruction $W = \{w_i\}_{i=1}^L$ with L words. The agent is required to interpret this instruction to navigate to the target location.

At step t on node V_t , the agent observes (1) a panorama $\mathcal{R}_t = \{r_{t,i}\}_{i=1}^K$ represented by K images from different views, $r_{t,i}$ is the extracted image feature of the i th view; (2) neighboring navigable nodes $\mathcal{N}(V_t)$, where $\mathcal{N}(V_t) \subset \mathcal{V}$; (3) orientations and coordinates of these neighboring nodes. The agent should choose a neighboring node to step to or stop at current location. Navigation is considered successful if the agent stops within 3 meters of the target.

We treat navigation as a process of searching temporary target to navigate towards next among unvisited nodes on \mathcal{G}_t . A temporary target can be either a local neighboring node or remote unvisited node. By adding a virtual stop node connected to all nodes on the graph, we can also model the stop action. We refer fidelity trajectory to a path from initial node to an unvisited node *without detours*.

3.2 Model Overview

The overall framework is shown in Fig. 3(a). Our planning method consists of three components: (1) We construct a directed graph for explored area. Considering the directional nature of navigation, we propose an Orientation-aware Panorama Encoder(OPE) to extract orientation-aware vision features for edges. Edge sequence is used to represent directed trajectory and align with the instruction. (2) We maintain a path (*i.e.* directed fidelity trajectory) to each unvisited node and assess the alignment between each path and the instruction at each step. Only few neighboring nodes are added at each step, we only compute for these relevant new paths. Since each node corresponds to a path, we calculate a path embedding with a matching assessment module(MAM) that encodes the instruction-path alignment and stores the embedding on the node. (3) We compare the path embeddings of all unvisited nodes(candidates) with candidate comparison module(CCM)(shown in Fig. 3(c)) to determine which aligns best with the instruction. We select the best aligned candidate node as the temporary target. Then the agent can navigate to the temporary target along the shortest path on constructed graph. Since the path embedding is an estimate of the alignment between instruction and a path to the node, CCM ensures that the agent navigates by fully considering the instruction-trajectory alignment.

3.3 Orientation-Aware Directed Graph Construction

We construct a directed graph to support path representation, incrementally adding nodes and edges as exploration proceeds. Previous methods [2, 8, 44] represent visited nodes with panorama features. However, panoramas of unvisited nodes are inaccessible as the agent has not visited them. So they represent these nodes with views towards them, leading to inconsistent representation between visited and unvisited nodes. In contrast, we extract visual features for the directed edges on the graph. These edge features represent the visual observations

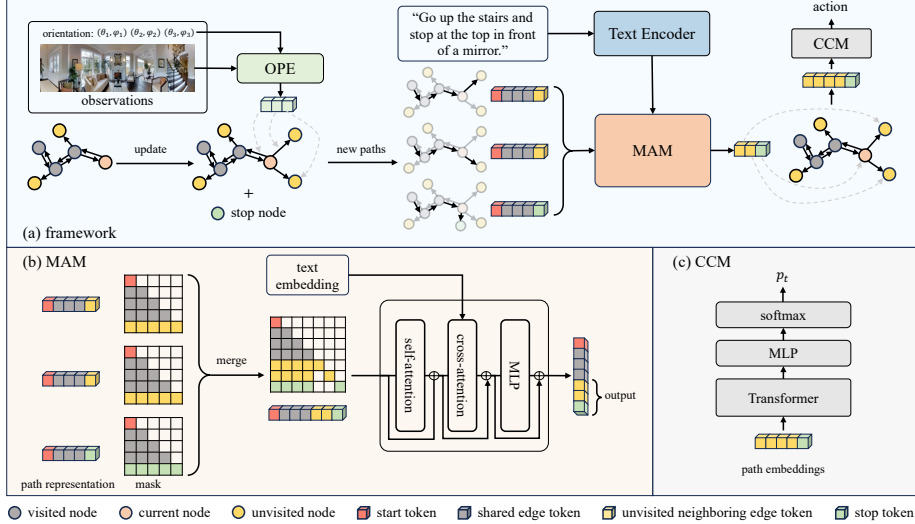


Fig. 3: Illustration of our model. (a) is the overall framework of our method. At each step, we update the graph, extract path embeddings, and predict actions. (b) depicts the matching assessment module(MAM). Each token is an edge feature. We compute path embeddings for each newly observed nodes with cross-modal transformer and impose a causal mask to reduce computational cost. (c) shows the candidate comparison module(CCM). We gather path embeddings of unvisited nodes and forward them into a single layer transformer followed by a MLP to predict temporary target.

when facing specific directions from each node. As edges rely only on observable views rather than node panoramas, they enable consistent representation for both visited and unvisited nodes.

Let $\mathcal{G}_t = \{\mathcal{V}_t, \mathcal{E}_t\}$ be the directed graph at step t . Neighboring nodes $\mathcal{N}(V_t)$ are observed, and some nodes in $\mathcal{N}(V_t)$ have already existed in \mathcal{G}_t while others are newly observed and added to \mathcal{G}_t . Besides, directed edges from V_t to nodes in $\mathcal{N}(V_t)$ together with features extracted by orientation-aware panorama encoder(OPE) will also be added in \mathcal{G}_t , as shown in Fig. 3(a).

Visual Representation For Edges. We extract orientation-aware panorama feature by using relative orientation as query to attend panorama features in cross-attention manner, as shown in the OPE module in Fig. 3(a). Assume that (ϕ, θ) represents the relative heading and elevation of a specific edge when compared to the agent’s current orientation. To extract feature for the edge, we first encode (ϕ, θ) as follows:

$$x^a = [\sin(\phi), \cos(\phi), \sin(\theta), \cos(\theta)] W^a, \quad (1)$$

where $W^a \in \mathbb{R}^{4 \times d}$ are learnable weights and d is the dimension of feature. Then we denote $X_t^a = \{x_{t,i}^a\}_{i=1}^{|\mathcal{N}(V_t)|}$ as orientation features of adjacent edges of V_t .

We then encode the panorama feature. As the panorama $\mathcal{R}_t = \{r_{t,i}\}_{i=1}^K$ is represented by K images from different views and each view corresponds to a fixed orientation, we can also get the relative heading and elevation of each view. To encode the panorama, we first concatenate each view feature $r_{t,i}$ with corresponding relative orientation $(\phi_{t,i}, \theta_{t,i})$ and then map it to dimension d via a linear projection:

$$x_{t,i}^p = [r_{t,i}; \sin(\phi_{t,i}), \cos(\phi_{t,i}), \sin(\theta_{t,i}), \cos(\theta_{t,i})]W^p. \quad (2)$$

Then we denote $X_t^p = \{x_{t,i}^p\}_{i=1}^K$ as the panoramic view representation. By considering views as patches of the panorama and orientations as position embedding, X_t^p is like the encoded ViT [12] input.

We adopt transformer decoder [43] to extract orientation-aware panorama feature:

$$E_t = \text{TransformerDecoder}(X_t^a, X_t^p), \quad (3)$$

where $E_t = \{e_{t,i}\}_{i=1}^{|\mathcal{N}(V_t)|}$ are extracted edge features. By taking X_t^a as the query input and X_t^p as the key-value input, panorama is attended by orientation in the first layer. In the subsequent layers, both orientation and vision information are used to query relevant views in the panorama. Therefore, each directed edge only focuses on a specific region of the panorama, thus orientation-aware panorama feature is extracted.

3.4 Planning with Fidelity Trajectory

We maintain a directed fidelity trajectory with a stack during navigation for each unvisited node on \mathcal{G}_t . When a new node is added, it is pushed into the stack. When the agent needs to backtrack, the top node is popped off the stack until a different node than the top is reached. In this way, detours are removed by popping nodes off the stack. With the maintained trajectory of each node, we propose MAM to assess its alignment with instruction. We then determine an unvisited node on \mathcal{G}_t to navigate next by comparing this alignment with CCM.

Matching Assessment Module(MAM). MAM is a multi-layer transformer decoder that extracts a path embedding for each path, which estimates how much the trajectory match with the instruction. These path embeddings are then stored on the corresponding nodes in graph \mathcal{G}_t .

For a single path with length l , we represent the path with the edge features along it, denoted as $X^h = [e_1, e_2, \dots, e_l]$. $\{e_i\}_{i=1}^l$ are the edge features on the path. Then we add it with position embedding and forward it together with text embedding X^w into MAM like follows:

$$\begin{aligned} X^{h'} &= X^h + P_l, \\ X^o &= \text{TransformerDecoder}(X^{h'}, X^w), \end{aligned} \quad (4)$$

where P_l is a matrix including l position embeddings [43] and X^w is the text embedding extracted by a text encoder, which is a multi-layer transformer [43].

By utilizing two sequences from two modalities as inputs, the transformer is capable of effectively estimating their alignment, as extensively demonstrated in Vision and Language Pretraining studies [9, 22, 25]. The last token of the encoded sequence X^o provides the path embedding, which is an estimation of the text-path alignment.

To support STOP action, we add a stop node that connected to all nodes in \mathcal{G}_t . Since we represent a path using edge features, we assign a learnable stop token, initialized with a zero vector, to all edges that point to the stop node. The path embedding corresponds to the stop node is stored on current node. So the agent decides to stop at current location when it find that the path up to here appending with stop token matches instruction the best. Here stop token serves as the End-of-Sentence token in text generation. For each path, we also add a learnable start token that serves as the Start-of-Sentence token.

At step t , considering the stop node, assume there are N neighboring unvisited nodes to be updated. Computing path embeddings separately for these nodes requires multiple forward passes, which is inefficient. *Actually, we can compute these embeddings in a single forward pass*, as shown in Fig. 3(b). Note that these paths share the same prefix from the initial node to the current node. We can avoid duplicated computation by imposing a causal mask in the self-attention block. The causal mask is a lower triangular matrix that only allows a token to attend to its previous tokens. We merge these causal masks together and forward all relevant tokens into the transformer block. The merged mask still satisfies the constraint that each token can only attend to its previous tokens. We further reduce the computational cost by adopting the KV-cache technique [35], which stores the shared prefix tokens to avoid unnecessary re-calculation. In this way, we can equivalently compute the path embeddings for multiple newly observed nodes in a single forward pass. The last several output tokens (colored yellow and green) are path embeddings and used to update the graph, *i.e.*, stored on corresponding graph nodes.

Candidate Comparison Module(CCM). CCM compares the path embeddings from candidate nodes to predict a temporary target. As each path embedding encodes the alignment between a trajectory and the instruction, the CCM actually selects a path that is most aligned with the instruction. In this way, we make navigation planning fully according to the instruction-trajectory alignment. As shown in Fig. 3(c), we collect path embeddings X_t^e and input them into CCM. CCM consists of a single-layer transformer encoder [43] and a MLP. The transformer encoder is responsible for the comparison of the candidates and the MLP maps the encoded path embedding to a 1D score:

$$\begin{aligned} X_t^{e'} &= \text{TransformerLayer}(X_t^e), \\ s_t &= \text{MLP}(X_t^{e'}), \\ p_t &= \text{softmax}(s_t). \end{aligned} \tag{5}$$

Score s_t reflects the relative text-path alignment after comparison. An alternative way is removing the transformer and directly compute the alignment score s_t .

However, this method assesses each path independently without any comparison, which can make the decision-making process more challenging and potentially result in a decrease in performance(demonstrated in Sec. 4). It is then normalized with softmax and we select the highest score node as the temporary target. The agent will navigate to it along the shortest path on \mathcal{G}_t . If the stop node is selected, the agent believes the instruction is completed and stops the navigation.

3.5 Pretraining and Fine-tuning

Pretraining. Pretraining is proved to be helpful in previous works [2, 7, 17, 33]. PREVALENT [17] synthesized a substantial amount of data for pretraining, and we utilized the same data as their work. We use Masked Language Modeling(MLM) [11] for pretraining. Concretely, we randomly replace 15% of the input tokens with a special [MASK] token and forward these masked tokens into the text encoder. We also extract a sequence of edge features to represent the path. Then the encoded text tokens and path representation are input into an additional transformer decoder. The corresponding outputs are used to predict the masked tokens. MLM helps the model learn aligned text and path representations. In VLN, a key difference from other vision-and-language tasks is that we align a text sequence with a sequence of vision representations rather than a single image.

Fine-tuning. We train the agent with a mixture of teacher-forcing and student-forcing as previous methods [2,8]. In the teacher-forcing stage, the agent navigate according the ground truth actions a_t^* and the loss is calculated as follows:

$$\mathcal{L}_{TF} = \frac{1}{T_{TF}} \sum_{t=1}^{T_{TF}} \text{CE}(p_t, a_t^*), \tag{6}$$

CE is the cross entropy loss and T_{TF} is the number of navigate steps. In the student-forcing stage, we sample an action from the distribution predicted by the agent so that the agent can explore the environment and reduce the exposure bias. We train the agent with heuristic pseudo label a_t^{pseudo} :

$$\mathcal{L}_{SF} = \frac{1}{T_{SF}} \sum_{t=1}^{T_{SF}} \text{CE}(p_t, a_t^{pseudo}). \tag{7}$$

When the agent deviate from ground truth path, we take the nearest node on ground truth path as the pseudo label to encourage agent to learn backtracking strategy. If there is no unvisited node on ground truth path, we take the nearest node on the shortest path from current node to target node as the pseudo label. The agent navigate two times to compute \mathcal{L}_{TF} and \mathcal{L}_{SF} . The total loss is the weighted sum of them:

$$\mathcal{L} = \lambda \mathcal{L}_{TF} + (1 - \lambda) \mathcal{L}_{SF}, \tag{8}$$

where $\lambda \in (0, 1)$ is the weight.

4 Experiments

4.1 Datasets and Evaluation Metrics

R2R. Room-to-Room (R2R) dataset [4] contains 7,189 trajectories, with 3 instructions per trajectory, across 90 scenes. These scenes are divided into train, val unseen and test unseen splits with 61, 11 and 18 scenes respectively. A val seen split with the same scenes as the train split is also provided. All paths in R2R are the shortest paths between the start and target nodes.

RxR. Room-across-Room(RxR) [23] is a large multilingual dataset. The dataset has 126,000 instructions total, with 42,000 instructions in each of the three languages: English, Hindi, and Telugu. The paths in RxR are longer than those in R2R and are not the shortest possible routes. Additionally, the instructions in RxR are longer and contain more detailed descriptions compared to R2R.

Evaluation Metrics. We adopt the following evaluation metrics on R2R: (1) Trajectory length (TL): the average length of the agent’s path in meters. (2) Navigation error (NE): the average distance between the agent’s final position and the target location. (3) Success rate (SR): the ratio of successful navigations, where $NE < 3m$ is considered successful. (4) SR penalized by path length (SPL) [3]: TL longer than the shortest path is penalized. As paths in RxR dataset are shortest paths, TL and SPL are unsuitable. Instead, on RxR we use nDTW and sDTW [21] to measure trajectory similarity between the agent and ground truth. Please note that the fidelity trajectory that is removed detours is solely used for planning. It is not the agent’s actual trajectory used for evaluation.

4.2 Implementation Details

Module architecture. Text backbone is a 6 layer transformer encoder initialized with pretrained ALBEF [25]. For the multilingual RxR dataset, we use 12 layer mRoberta [10] instead. The layer numbers of the OPE, MAM, and CCM are 2, 4, and 1 respectively. All the hidden layer size is 768. Image features is extracted by DINOv2 [34]. We also report results with CLIP-ViT-base [37] feature for fair comparison.

Training details. On R2R dataset, we first pretrain PRET with learning rate $2e-5$ and batch size 16 on a single 24G 4090 GPU for 100,000 iterations(~ 5 hours). Then we fine-tune the model with learning rate $1e-5$ and batch size 8 on 4090 for 100,000 iterations(~ 25 hours). Optimizer is AdamW [30]. λ is set to 0.2. Augmented dataset [17] is used in both pretraining and fine-tuning. The best result is selected by SPL on val unseen split. On RxR dataset, we also follow the pretrain and fine-tune paradigm. Due to the longer instruction and trajectory, we use a smaller batch size 4. Marky [45] is used as augmented data. The best result is selected by sDTW on val unseen split.

Table 1: Comparison with other methods on R2R dataset. SPL is considered as the primary evaluation metric.

Methods	Val Seen				Val Unseen				Test Unseen			
	TL	NE↓	SR↑	SPL↑	TL	NE↓	SR↑	SPL↑	TL	NE↓	SR↑	SPL↑
Seq2Seq-SF [4]	11.33	6.01	39	-	8.39	7.81	22	-	8.13	7.85	28	18
Speaker-Follower [14]	-	3.36	66	-	-	6.62	35	-	14.82	6.62	35	28
RCM [46]	10.65	3.53	67	-	11.46	6.09	43	-	11.97	6.12	43	38
Regretful [32]	-	3.23	69	63	-	5.32	50	41	-	5.69	56	40
EnvDrop [40]	11.00	3.99	62	59	10.70	5.22	52	48	11.66	5.23	51	47
PREVALENT [17]	10.32	3.67	69	65	10.19	4.71	58	53	10.51	5.30	54	51
NvEM [1]	11.09	3.44	69	65	11.83	4.27	60	55	12.98	4.37	58	54
SSM [44]	14.70	3.10	71	62	20.70	4.32	62	45	20.40	4.57	61	46
RecBert [18]	11.13	2.90	72	68	12.01	3.93	63	57	12.35	4.09	63	57
HAMT [7]	11.15	2.51	76	72	11.46	2.29	66	61	12.27	3.93	65	60
MTVM [28]	-	2.67	74	69	-	3.73	66	59	-	3.85	65	59
DUET [8]	12.32	2.28	79	73	13.94	3.31	72	60	14.73	3.65	69	59
AZHP [16]	-	-	-	-	14.05	3.15	72	61	14.95	3.52	71	60
Meta-Explore [20]	11.95	2.11	81	75	13.09	3.22	72	62	14.25	3.57	71	61
GridMM [47]	-	-	-	-	13.27	2.83	75	64	14.43	3.35	73	62
BEVBert [2]	13.56	2.17	81	74	14.55	2.81	75	64	15.87	3.13	73	62
Ours(CLIP)	11.48	2.60	74	69	12.21	3.12	71	63	13.87	3.12	72	62
Ours(DINov2)	11.25	2.41	78	72	11.87	2.90	74	65	12.21	3.09	72	64

4.3 Results

Comparison on R2R. Tab. 1 compares our approach with previous methods. On the val unseen and test unseen split, our method achieves comparable performance with previous SOTA method BEVBert on the primary metric SPL, while our method significantly reduces the computational cost. Besides, BEVBert [2] adopts additional depth information to construct bird’s-eye view input. On the val seen split, our method performs sub-optimal as PRET does not take the environment layout as input, which means it less tends to overfit on the seen environment. Also, in the VLN literature, all visual features are pre-extracted and remains unchanged, the fixed feature may not contains appropriate feature for navigation. Therefore, we also investigate different vision features for VLN, as shown in Tab. 1. We find that DINOv2 [34] feature is better than CLIP feature. DINOv2 is trained on large-scale curated data in self-supervised manner, which leads to more general purpose feature. Moreover, DINOv2 feature contains geometric information as it can be used to predict depth map according to [34]. Geometric information is shown helpful in [19].

Tab. 2 shows the parameter count, computational cost, and latency of different graph-based methods. We present GFLOPs of different navigation steps. Our method contains fewer parameters as we use a single stream model unlike [2, 8] use dual stream transformer. Besides, our decoder computational cost is only 19% of DUET and 3% of BEVBert at step 10. These suggests that our method is significantly more efficient than previous SOTA graph-based methods. As shown in Tab. 2, this results in lower training and inference latency.

Table 2: Comparison of latency and computational cost on R2R val unseen. * indicates we reproduce DUET with CLIP feature for fair comparison. When computing GFLOPs, we simplify the calculation by assuming that each vertex of the graph is connected to 4 adjacent vertices (average degree in R2R), and at each step, 3 new vertices are encountered. The computational cost of text encoder and visual encoder is omitted for fair comparison.

Methods	R2R val unseen				Latency(ms)		GFLOPs			Params
	TL	NE↓	SR↑	SPL↑	train	inference	1 step	10 steps	20 steps	
DUET* [8]	14.0	3.2	71.6	61.1	11.8	8.6	2.4	4.8	7.5	90M
BEVBert [2]	14.6	2.8	74.9	63.6	16.2	10.6	33.7	35.5	37.5	90M
Ours	12.2	3.1	71.0	62.7	7.7	4.1	0.6	0.9	1.2	64M

Table 3: Comparison with previous methods on RxR dataset. † indicates approaches that utilize additional augmented data. nDTW represents the normalized DTW distance between the agent’s trajectory and ground truth path. sDTW is the nDTW weighted by success rate.

Methods	Val Seen				Val Unseen			
	NE↓	SR↑	nDTW↑	sDTW↑	NE↓	SR↑	nDTW↑	sDTW↑
LSTM [23]	10.7	25.2	42.2	20.7	10.9	22.8	38.9	18.2
EnvDrop+ [39]	-	-	-	-	-	43.6	55.7	-
HAMT [7]	-	59.4	65.3	50.9	-	56.5	63.1	48.3
EnvEdit [24]	-	67.2	71.1	58.5	-	62.8	68.5	54.6
MPM [13]	-	67.7	71.0	58.9	-	63.5	67.7	54.5
MARVEL [45]†	3.0	75.9	79.1	68.8	4.5	64.8	70.8	57.5
BEVBert [2]†	3.2	75.0	76.3	66.7	4.0	68.5	69.6	58.6
Ours(CLIP)†	2.6	77.0	78.0	67.5	3.3	71.2	71.8	60.4
Ours(DINOv2)†	2.4	79.3	80.4	70.7	3.2	72.8	73.4	62.4

Comparison on RxR. Tab. 3 reports the results on the RxR dataset. RxR is challenging as instructions and paths in RxR are longer than R2R. While the alignment between path and detailed path description is what our method skills at, our method outperforms previous methods among all metrics at a lower computational cost. PRET achieves a 1.8% improvement on the main metric sDTW on the val unseen split, demonstrating that our model better understands and follows instructions. This result highlights the advantage of our planning with trajectory strategy, which considers the alignment between instructions and paths and enables efficient decision-making in the global space. Additionally, with DINOv2 features, the result is further improved.

4.4 Ablation Study

We conducted ablation experiments on various components of PRET, including the directed path representation and module ablation. Results are reported on the R2R val unseen split.

Table 4: Comparison of undirected and directed path representation.

Methods	TL	NE↓	SR↑	SPL↑
undirected	15.62	3.59	68.28	56.77
directed	11.87	2.90	73.78	65.16

Table 5: Ablation study on modules.

Methods	TL	NE↓	SR↑	SPL↑
1 MAM	12.04	3.99	62.32	54.48
2 MAM+CCM	12.15	3.54	65.94	57.32
3 MAM+OPE	12.18	3.15	71.60	63.07
4 MAM+OPE+CCM	11.87	2.90	73.78	65.16



Fig. 4: Comparison of orientation panoramic view and single candidate view.

Directionality in Path Representation. We compare whether to incorporate directionality for path representations in Tab. 4. Undirected path is represented by a sequence of node features. The node features is extracted by forwarding panoramic views into a 2-layer transformer encoder and averages the output tokens. Directed path is represented by a sequence of orientation-aware panorama features. According to the results, directed path representation outperforms undirected node features among all metrics. Specifically, it gains 8.39% improvement on SPL and 5.5% on SR. This indicates that edge feature is more suitable to represent the directional navigation process and align with the instruction. Node features does not distinguish different path directions and provide redundant information in a path, which hinders the alignment.

Module ablation. Tab. 5 shows ablation results on modules. Row 4 is the full model introduced in Sec. 3. OPE extracts orientation-aware features for edges, MAM extracts path embeddings while CCM compares them to select the temporary target. When removing CCM (Row 3), we only predict a single alignment score rather than a path embedding for each path without comparison. The score is then normalized and used to predict the temporary goal. Performance drops 2.09% in this case, as selecting a path based solely on a single score, without comparing different paths, is more difficult.

In row 2, we remove OPE, panorama features are not used and only a single 60-degree view towards each neighboring node is adopted as edge features. Performance drops 7.84% as the limited field-of-view cannot fully represent the path. The trajectory represented by these views is discrete and does not form a continuous change of views. In some cases, a single view towards candidates provides no information for navigation, as shown in Fig. 4. When go up the stairs, the upper view only sees the wall and does not help the navigation at all. While using the panorama, the agent can dynamically see a larger region. These results shows our graph representation with panorama encoder provides rich representation for trajectories. Row 1 demonstrates the performance when using only MAM. The performance is poor without other model components, demonstrating the effectiveness of these modules.

4.5 Visualization

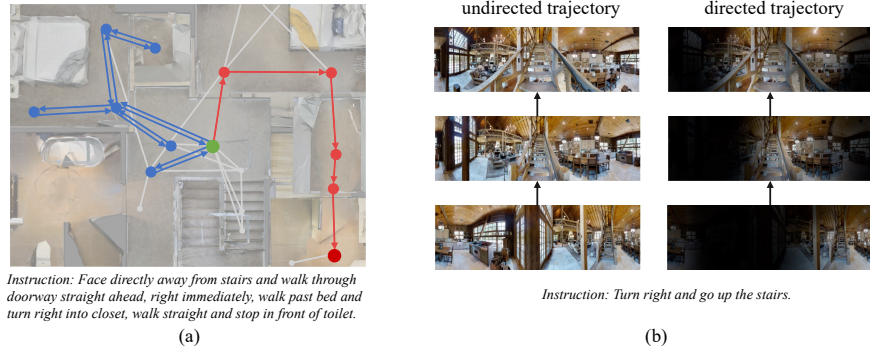


Fig. 5: (a) Visualization of the agent’s navigation process, showcasing its ability to learn a backtracking strategy. (b) Visualizing attention weights in OPE to illustrate the distinction between undirected and directed trajectory representations.

Fig. 5(a) shows a qualitative example of our model’s behavior. Initially placed at the green starting node, the agent wrongly decides to explore the upper left area (blue path). It takes many unnecessary detours there, but through comparing the trajectory alignment with the instructions, realizes these paths do not match the text well. Therefore, it backtracks to the start and correctly navigates to the destination (red path). This demonstrates the capability of PRET to learn complex backtracking behaviors.

Fig. 5(b) shows the difference between undirected and directed trajectory representations. The former represents a path using panorama features stored on nodes, which eliminates directional discrepancies since the same node provides the same feature regardless of orientation, and contains redundant information compared to “stairs” in the instruction. In contrast, our proposed approach represents a path using orientation-aware panorama features on edges. This eliminates redundant information and focuses on the “stairs” better captures the directional navigation process.

5 Conclusion

In this work, we present an efficient method that planning with directed fidelity trajectory, explicitly leveraging the alignment between instructions and trajectories for navigation. Additionally, we consider the directional nature of navigation and introduce a directed graph construction during navigation, storing vision information on directed edges. This approach provides rich directed path representation and enhances instruction-trajectory alignment. Experiments demonstrate that our method achieves strong performance while being significantly more efficient than previous methods. Our method has certain limitations that we need to address. For example, navigation requires environment layout information in some cases. We recognize the need to investigate potential solutions to overcome these challenges.

Acknowledgments

This work was supported partially by the National Key Research and Development Program of China (2023YFA1008503), NSFC(U21A20471, 62206315), Guangdong NSF Project (No. 2023B1515040025, 2020B1515120085, 2024A1515-010101), Guangzhou Basic and Applied Basic Research Scheme(2024A04J4067).

References

1. An, D., Qi, Y., Huang, Y., Wu, Q., Wang, L., Tan, T.: Neighbor-view enhanced model for vision and language navigation. In: ACM MM (2021)
2. An, D., Qi, Y., Li, Y., Huang, Y., Wang, L., Tan, T., Shao, J.: Bevbort: Multimodal map pre-training for language-guided navigation. In: ICCV (2023)
3. Anderson, P., Chang, A.X., Chaplot, D.S., Dosovitskiy, A., Gupta, S., Koltun, V., Kosecka, J., Malik, J., Mottaghi, R., Savva, M., Zamir, A.R.: On evaluation of embodied navigation agents. CoRR (2018)
4. Anderson, P., Wu, Q., Teney, D., Bruce, J., Johnson, M., Sünderhauf, N., Reid, I., Gould, S., Van Den Hengel, A.: Vision-and-language navigation: Interpreting visually-grounded navigation instructions in real environments. In: CVPR (2018)
5. Chaplot, D.S., Gandhi, D., Gupta, A., Salakhutdinov, R.: Object goal navigation using goal-oriented semantic exploration. In: NeurIPS (2020)
6. Chaplot, D.S., Gandhi, D., Gupta, S., Gupta, A., Salakhutdinov, R.: Learning to explore using active neural SLAM. In: ICLR (2020)
7. Chen, S., Guhur, P.L., Schmid, C., Laptev, I.: History aware multimodal transformer for vision-and-language navigation. NeurIPS (2021)
8. Chen, S., Guhur, P., Tapaswi, M., Schmid, C., Laptev, I.: Think global, act local: Dual-scale graph transformer for vision-and-language navigation. In: CVPR (2022)
9. Chen, Y.C., Li, L., Yu, L., El Kholy, A., Ahmed, F., Gan, Z., Cheng, Y., Liu, J.: Uniter: Universal image-text representation learning. In: ECCV (2020)
10. Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., Stoyanov, V.: Unsupervised cross-lingual representation learning at scale. In: ACL (2020)
11. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: BERT: Pre-training of deep bidirectional transformers for language understanding. In: NAACL (2019)
12. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N.: An image is worth 16x16 words: Transformers for image recognition at scale. In: ICLR (2021)
13. Dou, Z., Gao, F., Peng, N.: Masked path modeling for vision-and-language navigation. CoRR (2023)
14. Fried, D., Hu, R., Cirik, V., Rohrbach, A., Andreas, J., Morency, L.P., Berg-Kirkpatrick, T., Saenko, K., Klein, D., Darrell, T.: Speaker-follower models for vision-and-language navigation. In: NeurIPS (2018)
15. Fuentes-Pacheco, J., Ascencio, J.R., Rendón-Mancha, J.M.: Visual simultaneous localization and mapping: a survey. Artif. Intell. Rev. (2015)
16. Gao, C., Peng, X., Yan, M., Wang, H., Yang, L., Ren, H., Li, H., Liu, S.: Adaptive zone-aware hierarchical planner for vision-language navigation. In: CVPR (2023)
17. Hao, W., Li, C., Li, X., Carin, L., Gao, J.: Towards learning a generic agent for vision-and-language navigation via pre-training. In: CVPR (2020)

18. Hong, Y., Wu, Q., Qi, Y., Rodriguez-Opazo, C., Gould, S.: Vln bert: A recurrent vision-and-language bert for navigation. In: CVPR. pp. 1643–1653 (2021)
19. Huo, J., Sun, Q., Jiang, B., Lin, H., Fu, Y.: Geovln: Learning geometry-enhanced visual representation with slot attention for vision-and-language navigation. In: CVPR (2023)
20. Hwang, M., Jeong, J., Kim, M., Oh, Y., Oh, S.: Meta-explore: Exploratory hierarchical vision-and-language navigation using scene object spectrum grounding. In: CVPR (2023)
21. Ilharco, G., Jain, V., Ku, A., Ie, E., Baldrige, J.: General evaluation for instruction conditioned navigation using dynamic time warping. In: NeurIPS Workshop (2019)
22. Kim, W., Son, B., Kim, I.: Vilt: Vision-and-language transformer without convolution or region supervision. In: ICML (2021)
23. Ku, A., Anderson, P., Patel, R., Ie, E., Baldrige, J.: Room-across-room: Multilingual vision-and-language navigation with dense spatiotemporal grounding. In: EMNLP (2020)
24. Li, J., Tan, H., Bansal, M.: Envedit: Environment editing for vision-and-language navigation. In: CVPR (2022)
25. Li, J., Selvaraju, R., Gotmare, A., Joty, S., Xiong, C., Hoi, S.C.H.: Align before fuse: Vision and language representation learning with momentum distillation. NeurIPS (2021)
26. Li, X., Li, C., Xia, Q., Bisk, Y., Celikyilmaz, A., Gao, J., Smith, N.A., Choi, Y.: Robust navigation with language pretraining and stochastic sampling. In: EMNLP-IJCNLP (2019)
27. Li, X., Li, C., Xia, Q., Bisk, Y., Celikyilmaz, A., Gao, J., Smith, N.A., Choi, Y.: Robust navigation with language pretraining and stochastic sampling. In: EMNLP-IJCNLP (2019)
28. Lin, C., Jiang, Y., Cai, J., Qu, L., Haffari, G., Yuan, Z.: Multimodal transformer with variable-length memory for vision-and-language navigation. In: ECCV (2022)
29. Liu, C., Zhu, F., Chang, X., Liang, X., Ge, Z., Shen, Y.: Vision-language navigation with random environmental mixup. In: ICCV (2021)
30. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. In: ICLR (2019)
31. Ma, C.Y., Lu, J., Wu, Z., AlRegib, G., Kira, Z., Socher, R., Xiong, C.: Self-monitoring navigation agent via auxiliary progress estimation. In: ICLR (2019)
32. Ma, C., Wu, Z., AlRegib, G., Xiong, C., Kira, Z.: The regretful agent: Heuristic-aided navigation through progress estimation. In: CVPR (2019)
33. Majumdar, A., Shrivastava, A., Lee, S., Anderson, P., Parikh, D., Batra, D.: Improving vision-and-language navigation with image-text pairs from the web. In: ECCV (2020)
34. Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., Assran, M., Ballas, N., Galuba, W., Howes, R., Huang, P., Li, S., Misra, I., Rabbat, M.G., Sharma, V., Synnaeve, G., Xu, H., Jégou, H., Mairal, J., Labatut, P., Joulin, A., Bojanowski, P.: Dinov2: Learning robust visual features without supervision. CoRR (2023)
35. Pope, R., Douglas, S., Chowdhery, A., Devlin, J., Bradbury, J., Levskaya, A., Heek, J., Xiao, K., Agrawal, S., Dean, J.: Efficiently scaling transformer inference. CoRR (2022)
36. Qi, Y., Wu, Q., Anderson, P., Wang, X., Wang, W.Y., Shen, C., Hengel, A.v.d.: Reverie: Remote embodied visual referring expression in real indoor environments. In: CVPR (2020)

37. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: ICML (2021)
38. Sennrich, R., Haddow, B., Birch, A.: Improving neural machine translation models with monolingual data. In: ACL (2016)
39. Shen, S., Li, L.H., Tan, H., Bansal, M., Rohrbach, A., Chang, K., Yao, Z., Keutzer, K.: How much can CLIP benefit vision-and-language tasks? In: ICLR (2022)
40. Tan, H., Yu, L., Bansal, M.: Learning to navigate unseen environments: Back translation with environmental dropout. In: NAACL (2019)
41. Thomason, J., Murray, M., Cakmak, M., Zettlemoyer, L.: Vision-and-dialog navigation. In: CoRL (2020)
42. Thrun, S.: Learning metric-topological maps for indoor mobile robot navigation. *Artif. Intell.* (1998)
43. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L.u., Polosukhin, I.: Attention is all you need. In: NeurIPS (2017)
44. Wang, H., Wang, W., Liang, W., Xiong, C., Shen, J.: Structured scene memory for vision-language navigation. In: CVPR (2021)
45. Wang, S., Montgomery, C., Orbay, J., Birodkar, V., Faust, A., Gur, I., Jaques, N., Waters, A., Baldridge, J., Anderson, P.: Less is more: Generating grounded navigation instructions from landmarks. In: CVPR (2022)
46. Wang, X., Huang, Q., Celikyilmaz, A., Gao, J., Shen, D., Wang, Y.F., Wang, W.Y., Zhang, L.: Reinforced cross-modal matching and self-supervised imitation learning for vision-language navigation. In: CVPR (2019)
47. Wang, Z., Li, X., Yang, J., Liu, Y., Jiang, S.: Gridmm: Grid memory map for vision-and-language navigation. In: ICCV (2023)
48. Zhu, F., Zhu, Y., Chang, X., Liang, X.: Vision-language navigation with self-supervised auxiliary reasoning tasks. In: CVPR (2020)