FILD-Nav:Vision-and-Language Navigation with Instruction Landmark Features in Continuous Environments

Chuangye Hu[1,2] , **Lulu Liu**[1] , **Huaiwei Si**[3] , **Yawen Zhao**[1] and **Nan Ding**[1,*]
[1]School of Computer Science and Technology,Dalian University of Technology
[2]College of Computer Science and Technology,Xinjiang Normal University
[3]Geely University of China
dhcy@mail.dlut.edu.cn, liululu655655@mail.dlut.edu.cn, huaiwei.si@geely.com,
zhaoyw@mail.dlut.edu.cn, dingnan@dlut.edu.cn

Abstract

Vision-and-language navigation (VLN) requires agents to follow natural language instructions to navigate autonomously in continuous environments. However, existing approaches often lack high-level semantic guidance in waypoint prediction and explicit language–landmark alignment in cross-modal planning. To address these limitations, we propose FILD-Nav, a vision-and-language navigation framework that integrates instruction landmark features. FILD-Nav extracts task-relevant landmarks from instructions and incorporates landmark semantics into both waypoint prediction and topological planning. Specifically, landmark-guided waypoint prediction improves waypoint relevance, while landmark-enhanced cross-modal planning enables more effective long-horizon navigation. Extensive experiments on the VLN-CE benchmark demonstrate that FILD-Nav consistently outperforms prior methods, achieving improvements of 2% in Success Rate (SR), 3% in Success weighted by Path Length (SPL), and 7% in Oracle Success Rate (OSR), particularly in unseen environments.

1 Introduction

Vision-and-Language Navigation (VLN) requires an embodied agent to follow natural language instructions to reach a target location in a visual environment[Anderson *et al.*, 2018]. While early VLN research focused on discrete graph-based settings[Wu *et al.*, 2024a], recent efforts have shifted toward navigation in continuous environments (VLN-CE)[Krantz *et al.*, 2020], which more closely reflect real-world embodied navigation scenarios. Operating in continuous spaces introduces additional challenges, including precise waypoint prediction, long-horizon planning, and robust alignment between linguistic instructions and visual observations.

A key aspect of human navigation is the use of landmarks. Humans naturally decompose navigation instructions into landmark-referenced subgoals (e.g., “turn left at the sofa” or “walk past the kitchen”), which provide semantic anchors that reduce ambiguity and facilitate planning. However, existing VLN-CE methods largely rely on implicit language representations or visual semantics extracted from observations, without explicitly modeling instruction-level landmark information[Tan *et al.*, 2024]. As a result, waypoint prediction and topological planning often lack direct grounding in task-relevant semantic cues, leading to suboptimal navigation decisions, especially in previously unseen environments.

Recent approaches have explored hierarchical planning, waypoint-based navigation, and semantic topological maps to address the complexity of VLN-CE[Zhou *et al.*, 2024; Shi *et al.*, 2025]. While these methods improve navigation efficiency, landmark information is typically derived from visual observations alone or treated as part of generic language embeddings. This design overlooks the fact that navigation instructions themselves provide rich, structured landmark information that can guide both local waypoint prediction and global planning. Moreover, without explicitly aligning instruction-derived landmarks with visual observations, agents may struggle to generalize across environments with varying layouts and appearances.

To address these limitations, we propose **FILD-Nav**, a landmark-aware navigation framework that **F**uses **I**nstruction-level **L**an**D**mark features into both waypoint prediction and topological planning. Our approach first extracts candidate landmarks from navigation instructions using a large language model, and then refines these landmarks through visual grounding during navigation. The resulting instruction-grounded landmark representations are used to guide waypoint prediction and to enhance node representations in a topological navigation graph, enabling more semantically aligned and robust decision-making.

We evaluate FILD-Nav on the VLN-CE benchmark and demonstrate consistent improvements over state-of-the-art methods. In particular, our method achieves significant gains in success rate, oracle success rate, and SPL on unseen environments, highlighting its strong generalization capability. Extensive ablation studies further validate the effectiveness of integrating instruction-level landmark features at different stages of the navigation pipeline.

In summary, our contributions are threefold:

- We introduce an instruction-level landmark representation for VLN-CE, explicitly extracting and grounding

*Corresponding author: Nan Ding

semantic landmarks from natural language instructions.

- We propose a landmark-aware waypoint prediction and topological planning framework that tightly aligns navigation decisions with task-relevant semantics.
- We demonstrate state-of-the-art performance on VLN-CE benchmarks, with particularly strong improvements in unseen environments.

2 Related Work

Vision-and-Language Navigation. VLN aims to enable embodied agents to follow natural language instructions to reach target locations. Early VLN methods were developed in discrete, graph-based environments, where navigation decisions are restricted to predefined viewpoints[Savva *et al.*, 2019; Chib and Singh, 2024]. Recent work has shifted toward VLN-CE, which require agents to operate in photorealistic 3D spaces with continuous actions[Hong *et al.*, 2022]. To address the increased complexity, prior approaches have explored imitation learning, reinforcement learning, and hierarchical navigation frameworks[Chen *et al.*, 2021; Wang *et al.*, 2024b]. Despite these advances, effectively grounding language instructions in continuous visual observations remains a central challenge.

Waypoint-based and Hierarchical Navigation. To improve navigation efficiency and long-horizon planning, several methods decompose navigation into hierarchical stages, typically involving waypoint prediction followed by low-level control[Krantz *et al.*, 2021]. Waypoint-based approaches reduce the action space and enable more stable navigation in continuous environments[Krantz and Lee, 2022; Wen *et al.*, 2024]. Some methods predict waypoints directly from visual observations[Raychaudhuri *et al.*, 2021], while others incorporate semantic cues or memory mechanisms to guide waypoint selection[Wu *et al.*, 2024b; Zheng *et al.*, 2025]. However, most existing approaches rely primarily on visual features or implicit language embeddings, without explicitly leveraging structured semantic information provided by navigation instructions.

Semantic and Topological Mapping. Topological maps provide an effective abstraction for long-term navigation by representing environments as graphs of navigable nodes[Chen *et al.*, 2021; Georgakis *et al.*, 2022]. Recent VLN-CE methods have incorporated semantic information into topological representations to improve planning and generalization[Chen *et al.*, 2022; Huang *et al.*, 2022]. These approaches typically extract semantic cues from visual observations and embed them into graph nodes[Liu *et al.*, 2024; Bao *et al.*, 2025]. While effective, such designs overlook instruction-level semantics and treat language as a global condition rather than a source of explicit, structured guidance for map construction and planning[Hong *et al.*, 2022; An *et al.*, 2024].

Language Grounding and Landmark-Based Navigation. Landmarks play a crucial role in human navigation and have been studied in both robotics and VLN literature[Lin *et al.*, 2024; Dai *et al.*, 2024]. Some prior works leverage object detections or scene labels as visual landmarks to assist navigation[Li *et al.*, 2023; He *et al.*, 2024]. Others use language attention mechanisms to implicitly align instructions with visual features[Wu *et al.*, 2024b; Asadi *et al.*, 2019]. In contrast to these methods, our approach explicitly extracts landmark information from navigation instructions and grounds it in visual observations during navigation. By incorporating instruction-grounded landmarks into both waypoint prediction and topological planning, our method achieves tighter alignment between language, perception, and planning.

Large Models for Vision-Language Tasks. Recent progress in large language models and vision-language models has significantly advanced multimodal reasoning capabilities[Cai *et al.*, 2024; Wang *et al.*, 2024a]. These models have been applied to various vision-language tasks, including navigation, for representation learning and semantic alignment[Bao *et al.*, 2025; Zhou *et al.*, 2024]. In this work, we leverage large models as auxiliary components to extract and refine instruction-level landmark representations[Qiao *et al.*, 2024; Xu *et al.*, 2025], rather than relying on them as end-to-end navigation policies. This design enables effective semantic grounding while maintaining a lightweight and interpretable navigation pipeline.

3 Methodology

3.1 Overview of FILD-Nav

To address the lack of high-level semantic understanding in waypoint prediction and the absence of explicit language–landmark alignment in cross-modal planning for continuous environments, we propose FILD-Nav, a landmark-aware VLN framework for continuous environments. As illustrated in Figure 1, FILD-Nav consists of three key modules: topological mapping, hierarchical planning, and low-level action control. The framework integrates large-model–driven instruction landmark extraction, landmark-aware waypoint prediction, and landmark-semantic–enhanced cross-modal planning, providing an effective solution for language-guided navigation in embodied agents.

3.2 Problem Formulation

We study Vision-and-Language Navigation in Continuous Environments, where an agent must follow a natural language instruction to navigate toward a target location without access to a predefined navigation graph. The task is instantiated in the Habitat simulator using panoramic RGB-D observations from Matterport3D and the R2R-CE benchmark.

At each episode, the agent starts from a given location and receives an instruction $\mathcal{W} = \{w_i\}_{i=1}^{L}$. The agent executes low-level actions from a discrete set: move forward 0.25m, turn left or right by 15°, and stop. An episode terminates when the agent issues the stop action or reaches the maximum time steps.

At time step t, the agent observes panoramic RGB and depth images $\mathcal{V}_t^{rgb} = \{v_{t,i}^{rgb}\}_{i=1}^{12}$ and $\mathcal{V}_t^{d} = \{v_{t,i}^{d}\}_{i=1}^{12}$, uniformly covering 360° with 30° intervals. Each view has a 90° field of view and resolution 256×256.

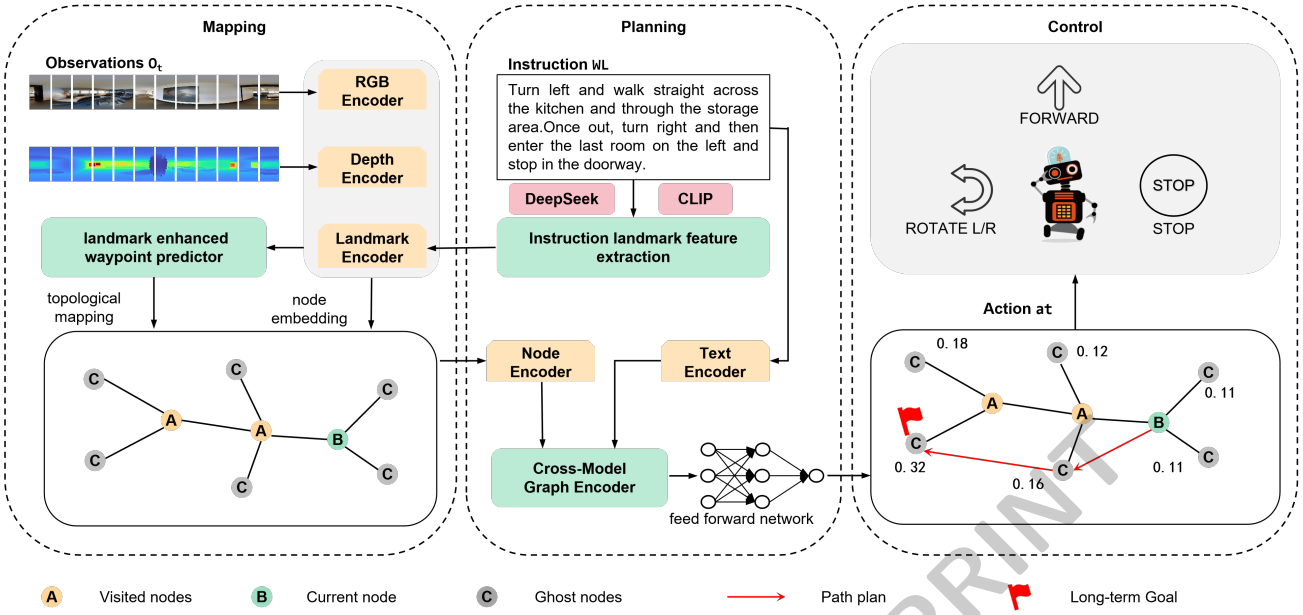

Figure 1: Overall Architecture of the Proposed FILD-Nav Framework.

3.3 Visually Grounded Instruction Landmark Features

As shown in Figure 2, the vision-constrained instruction landmark extraction framework consists of two stages: landmark generation and landmark correction, each built upon a different foundation model. In the first stage, we employ a large language model, DeepSeek, to extract ordered instruction landmarks and their co-occurring landmarks using customized prompts and in-context examples, activating task-relevant knowledge through text completion.We use DeepSeek as an off-the-shelf language model for offline instruction preprocessing; the navigation policy does not depend on proprietary APIs at inference time. These landmarks provide high-level semantic priors for navigation but may be inconsistent with actual visual observations. To address this issue, the second stage introduces a CLIP-based landmark correction module that leverages vision–language alignment to calibrate language-derived landmark priors using real-time visual evidence. By enforcing visual constraints on instruction landmarks, the proposed framework produces reliable landmark features that support downstream waypoint prediction and cross-modal planning. Natural language instructions contain rich landmark cues that are critical for both waypoint prediction and long-horizon path planning. Effective landmark extraction requires identifying navigation-ordered landmarks, as the order in which landmarks appear in language may not coincide with the order in which they should be searched during navigation. Moreover, some instruction-specified landmarks may not be directly observable, necessitating the incorporation of co-occurring landmarks derived from open-world spatial knowledge (e.g., beds are often associated with pillows or wardrobes in indoor environments). Accordingly, we design heuristic prompt templates that extract all concrete landmarks mentioned in the instruction, avoid abstract nouns, and reorder landmarks according to navigation logic.

Landmark extraction and co-occurrence generation are performed as an offline preprocessing step. Using the R2R-CE dataset as an example, we process 16,844 instructions with a customized DeepSeek-based pipeline, extracting 3–8 ordered primary landmarks per instruction and generating five co-occurring landmarks for each. These annotations are augmented into the dataset, enabling the agent to align a richer set of object-level semantics with visual observations.

However, without explicit visual grounding, language-derived landmark priors may be inconsistent with the actual environment and lead to erroneous guidance. To address this issue, we introduce a CLIP-driven landmark scoring module that dynamically challenges and refines landmark priors based on real-time visual evidence.

Vision-Constrained Landmark Scoring. Given an instruction I, a large language model first extracts an ordered set of primary landmarks U_{la} and their corresponding co-occurring landmarks U_{co}, forming the landmark set

$$U = \{U_{la}, U_{co}\}. \tag{1}$$

At step t, we extract visual feature F_O from the current observation via the CLIP image encoder. For adjacent primary landmarks U_{la}^{z} and U_{la}^{z+1}, their textual features F_z and F_{z+1} are obtained via the CLIP text encoder. The probability of U_{la}^{z} is computed as

$$P_z = \frac{\exp\left(\mathrm{sim}(F_O, F_z)/\tau\right)}{\exp\left(\mathrm{sim}(F_O, F_z)/\tau\right) + \exp\left(\mathrm{sim}(F_O, F_{z+1})/\tau\right)}, \tag{2}$$

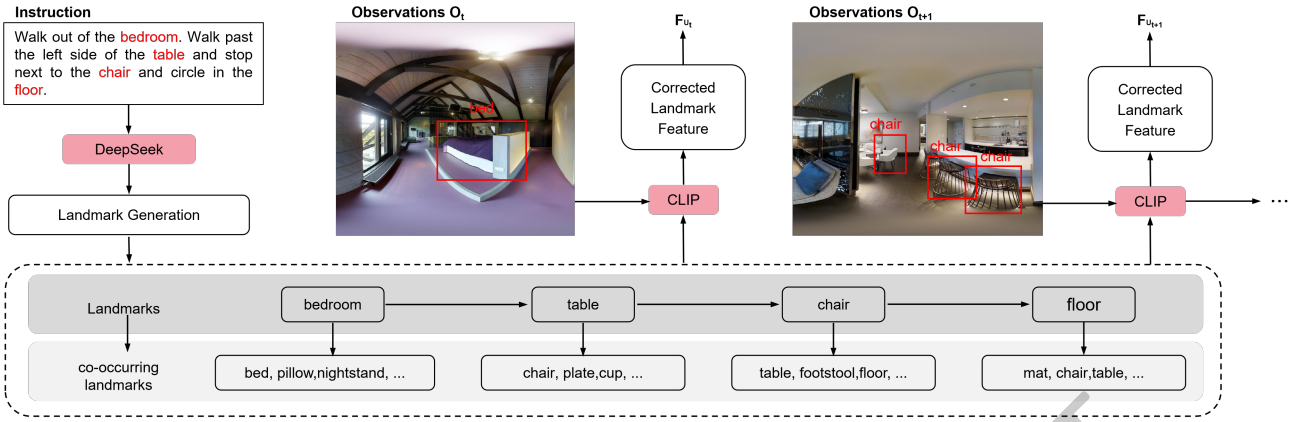

Figure 2: Vision-Constrained Instruction Landmark Extraction Framework.

where $\mathrm{sim}(\cdot)$ denotes cosine similarity and τ is a temperature parameter. We select the landmark with higher probability as the effective primary landmark.

After remembering the effective landmark U_{la}^{z} and its associated co-occurring landmark set U_{co}^{z}, we estimate the probability of observing U_{la}^{z} in each single-view observation O_n as

$$P_{U_{la}^{z}}^{(n)} = \frac{\exp(\mathrm{sim}(F_{U_{la}}, F_{O_n})/\tau)}{\sum_{m=1}^{K} \exp(\mathrm{sim}(F_{U_{la}}, F_{O_m})/\tau)}. \tag{3}$$

Similarly, the observation probabilities $P_{U_{co}^{z}}^{(n)}$ for each co-occurring landmark are computed.

To suppress noise introduced by landmark transitions and false discoveries, we introduce a learnable co-occurrence scoring module. Specifically, an embedding function $E_l(\cdot)$ consisting of a linear layer, ReLU activation, layer normalization, and dropout is applied to both instruction and observation features to obtain a state feature F_S. The co-occurring landmark score is defined as

$$S_{co} = \mathrm{sim}(F_S, F_{U_{co}}), \tag{4}$$

and the primary landmark score S_{la} is computed analogously.

The corrected landmark observation probability for view O_n is then given by

$$P_U^{(n)} = P_{U_{la}^{z}}^{(n)} \cdot S_{la} + \sum_{i=1}^{K} P_{U_{co_i}^{z}}^{(n)} \cdot S_{co_i}. \tag{5}$$

Finally, the visually corrected landmark feature is computed as

$$F_{U_t,n} = P_{U_{la}}^{t,n} \cdot S_{la} \cdot F_{U_{la}}^{t} + \sum_{i=1}^{N_{co}} P_{U_{co_i}}^{t,n} \cdot S_{co_i} \cdot F_{U_{co_i}}^{t}. \tag{6}$$

3.4 Waypoint Prediction with Instruction-Grounded Landmark Integration

To adapt the waypoint predictor to the introduced landmark features, we redesign and extend the original architecture. The enhanced waypoint predictor consists of an RGB visual encoder, a depth visual encoder, an instruction-grounded landmark feature encoder, a three-layer Transformer, and a decoupled linear classifier head. For visual feature extraction, a DINOv2 encoder pretrained on ImageNet is employed for RGB observations, while a ResNet-50 encoder pretrained on point-goal navigation tasks is used for depth images. Landmark features are obtained using the instruction landmark extraction method described in Section 3.3.

The Transformer comprises three layers, each with twelve self-attention heads, enabling effective modeling of spatial relationships across panoramic views. The classifier is implemented as a multilayer perceptron (MLP) that projects Transformer outputs into probabilities over spatially adjacent waypoint candidates. During training, the parameters of the visual encoders remain frozen, while the Transformer and classifier are optimized end-to-end.

At each time step, the agent receives RGB and depth observations, which are encoded separately to generate RGB features F_{RGB} and depth features F_{D}. Benefiting from skip connections within the encoders, these features capture both fine-grained low-level semantics (e.g., textures and colors) and geometric spatial information (e.g., navigability). Instruction-grounded landmark features F_U extracted by the large-model-based landmark module are incorporated as an additional semantic signal.

The three feature modalities are fused through a simple yet effective aggregation strategy. Specifically, average pooling and max pooling are applied to the RGB and landmark features to extract salient information. The pooled features are flattened, followed by a Dropout layer and a linear projection. Layer normalization is then applied to obtain the processed RGB feature F_{RGB} and landmark feature F_U. For depth features, average pooling is applied, followed by flattening and normalization to produce the processed depth fea-

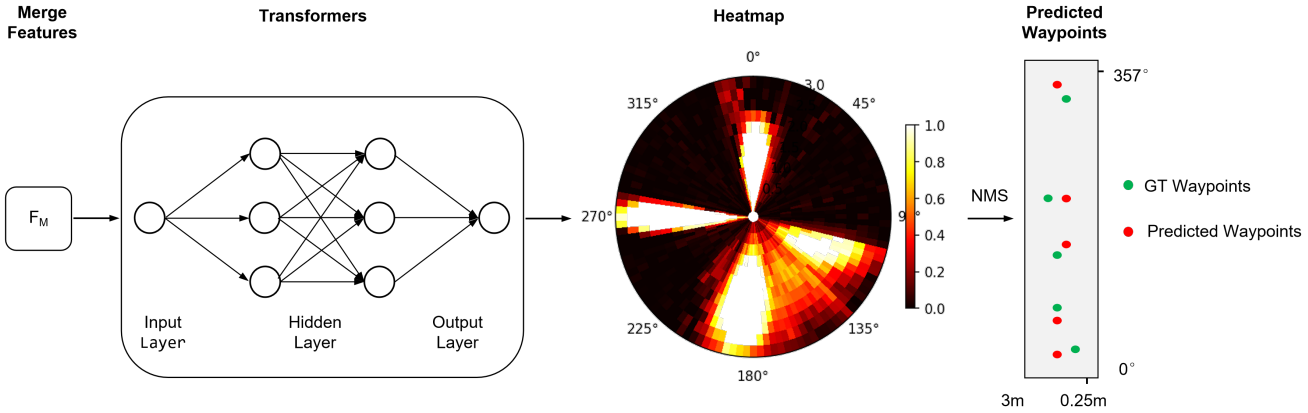

Figure 3: Landmark-Enhanced Waypoint Prediction.

ture F_{D}. The final fused feature F_M is computed as:

$$\begin{aligned} F_M = \mathrm{LN}(\mathrm{Dropout}(\mathrm{AP}(F_{\mathrm{RGB}}))) \\ + \mathrm{LN}(\mathrm{AP}(F_{\mathrm{D}})) \\ + \mathrm{LN}(\mathrm{Dropout}(\mathrm{MP}(F_U))). \end{aligned} \tag{7}$$

where $\mathrm{AP}(\cdot)$ and $\mathrm{MP}(\cdot)$ denote average pooling and max pooling, respectively, and $\mathrm{LN}(\cdot)$ denotes layer normalization.

The landmark-enhanced waypoint prediction process is shown in Figure 3. Each single-view image is square with a 90° field of view, covering three 30° angular sectors. Accordingly, we constrain the self-attention of each fused feature token to include neighboring sector features on both sides. The Transformer outputs tokens encoding sector-centered information, which are fed into the classifier to produce a panoramic heatmap. The heatmap has an angular resolution of 3° and a radial range from $0.25\,\mathrm{m}$ to $3.00\,\mathrm{m}$ with a step size of $0.25\,\mathrm{m}$, yielding a 120×12 spatial grid.

3.5 Landmark-Semantic Enhanced Cross-Modal Planning

The high-level planning module maps predicted waypoints into a dynamic topological graph augmented with landmark semantics and performs long-horizon path planning, producing a sequence of polar coordinates as navigation targets. The low-level control module executes primitive actions, including moving forward by $0.25\,\mathrm{m}$, turning left or right by 15°, and stopping. The navigation episode terminates once the agent reaches the target location or issues the stop action.

To support long-term planning, the agent incrementally constructs a topological map based on predicted waypoints. Waypoints that are visited or observed along the navigation trajectory are abstracted as nodes in a graph representation, denoted as

$$G_t = \langle \mathcal{N}_t, \mathcal{E}_t \rangle, \tag{8}$$

where $\mathcal{N}_t$ is the node set and $\mathcal{E}_t$ is the edge set. Each node $n_i \in \mathcal{N}_t$ encodes the panoramic visual observation and spatial location at that position. An edge $e_{i,j} \in \mathcal{E}_t$ is established if the corresponding node positions are directly reachable.

The topological graph is updated according to the spatial relationship between predicted waypoints and existing nodes. A waypoint localization function $F_L(\cdot)$ is employed to associate each predicted waypoint with a node in the graph. Specifically, F_L computes the Euclidean distance between the waypoint position and all existing nodes. If the minimum distance is smaller than a threshold γ, the corresponding node is returned as the localized node. This localization procedure is applied to every predicted waypoint to maintain a consistent topological representation.

As shown in Figure 4, the cross-modal planning module consists of a text encoder and a cross-modal graph encoder. Given the encoded topological map and the instruction representation as input, the module performs reasoning to predict a long-term goal node and outputs a topological path leading to the target.

For text encoding, we adopt BERT as the instruction encoder to extract contextualized word-level representations. Word embeddings are combined with positional and segment embeddings to form the instruction input, which is processed by stacked BertLayers to obtain instruction features with rich contextual semantics.

Each node embedding is enhanced with landmark-aware semantic features and global relative pose information, including its orientation and Euclidean distance relative to the current node. In addition, a navigation-step embedding encodes the most recent visitation time of each node, enabling the model to capture navigation history and temporal dependencies. Panoramic visual features augmented with landmark semantics are further combined with positional encodings to provide comprehensive node representations.

The encoded node representations and instruction features are jointly processed by a cross-modal graph encoder, following a mixed-input formulation similar to LXMERT. Within each encoder layer, we introduce a Graph-Aware Self-Attention (GASA) mechanism that explicitly incorporates topological structure:

$$\mathrm{GASA}(X) = \mathrm{Softmax}\left(\frac{XW_q(XW_k)^\top}{\sqrt{d}} + EW_e\right) XW_v, \tag{9}$$

where X denotes the node representations, E is the shortest-path distance matrix derived from the topological graph, and

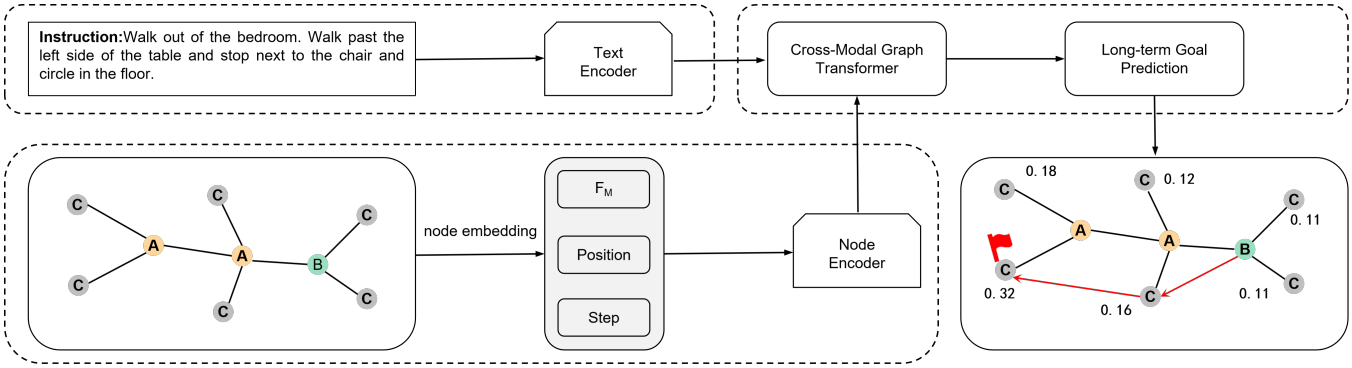

Figure 4: Cross-Modal Planning Module.

W_q, W_k, W_e, and W_v are learnable projection matrices.

The resulting cross-modal graph attention encodings capture fine-grained alignments between nodes, visual observations, and instruction semantics. A feed-forward network is applied to predict a stop score for each node. To prevent redundant exploration, nodes that have already been visited, as well as the current node, are masked during scoring. The agent selects an unvisited navigable node with the highest stop score as the long-term goal. To reach this goal, Dijkstra's algorithm is executed on the topological graph to compute the shortest path from the current node. If the selected node corresponds to the stop action, the agent terminates navigation at its current position.

4 Experiments

4.1 Experimental Setup

Dataset. The R2R-CE dataset is derived from the Room-to-Room (R2R) dataset [Anderson *et al.*, 2018] and adapted to continuous environments using the Habitat simulator [Savva *et al.*, 2019]. It is one of the most widely adopted benchmarks for VLN-CE. The dataset contains 16,844 path–instruction pairs collected from 90 real-world indoor scenes in the Matterport3D dataset, corresponding to 5,611 shortest-path trajectories. Each trajectory is annotated with approximately three English instructions, with an average path length of 9.89 meters and an average instruction length of 32 words.

The dataset is split into four subsets: training (*train*), seen validation (*val-seen*), unseen validation (*val-unseen*), and test. Following standard evaluation protocols, we report results on both validation splits. The *val-seen* split contains environments observed during training, while *val-unseen* consists of entirely novel environments, paths, and instructions, enabling an evaluation of generalization performance.

Based on R2R-CE, we further introduce large-model interfaces to preprocess all instructions by extracting instruction-grounded landmarks and generating co-occurring landmarks. These annotations are incorporated into the original dataset, resulting in an augmented version referred to as **R2R-CE with Instruction Landmarks**, which provides enriched semantic supervision for navigation.

Evaluation Metrics. An episode is considered successful if the agent executes the `STOP` action within 3 meters of the target location. We evaluate navigation performance using five standard metrics: **Trajectory Length (TL)**, which measures the average length of predicted trajectories; **Navigation Error (NE)**, defined as the average Euclidean distance between the final agent position and the target; **Success Rate (SR)**, indicating the percentage of successful episodes; **Oracle Success Rate (OSR)**, which considers the closest point to the goal along the trajectory; and **Success weighted by Path Length (SPL)**, which jointly evaluates navigation accuracy and efficiency.

Implementation Details. Our method is implemented using PyTorch and the Habitat simulator. All experiments are conducted on a Linux server equipped with an NVIDIA RTX 3090 GPU running Ubuntu 18.04. The model is trained for 20,000 iterations with a batch size of 50, and convergence is typically achieved after approximately four days. We employ the AdamW optimizer with a learning rate of 1×10^{-4}. Since the topological map is not available *a priori*, the agent interacts with the environment online to incrementally construct the map. Following prior work [An *et al.*, 2024; Radford *et al.*, 2021; Wijmans *et al.*, 2019], we initialize the model from a pre-trained checkpoint and continue training using scheduled sampling. The final model is selected based on the highest SPL achieved on the unseen validation split.

For visual encoding, multiple pre-trained models are adopted. A CLIP-pretrained ViT-B/32 model [Radford *et al.*, 2021] is used to extract global RGB features, while a ResNet-50 model pre-trained for point-goal navigation encodes depth observations. In addition, a CLIP-based learnable scoring module is employed to obtain instruction-grounded landmark features. The parameters of all pre-trained encoders are frozen during training. At each timestep, multimodal features are extracted by jointly processing the current instruction and visual observations.

4.2 Experimental Comparison

Table 1 compares FILD-Nav with state-of-the-art methods on the two validation splits of the R2R-CE dataset. The results show that our approach consistently outperforms existing methods in terms of OSR, SR, and SPL, while achieving performance comparable to the strong baseline ETPNav on trajectory length (TL) and navigation error (NE).

Method	Val Seen					Val Unseen					Test				
	TL↓	NE↓	OSR↑	SR↑	SPL↑	TL↓	NE↓	OSR↑	SR↑	SPL↑	TL↓	NE↓	OSR↑	SR↑	SPL↑
Seq2Seq [Krantz *et al.*, 2020]	9.26	7.12	46	37	35	8.64	7.37	40	32	30	8.64	7.37	40	32	30
LAW [Raychaudhuri *et al.*, 2021]	9.34	6.35	49	40	37	8.89	6.83	44	35	31	–	–	–	–	–
CMTP [Chen *et al.*, 2021]	–	7.10	56	36	31	–	7.90	38	26	23	–	–	–	–	–
HPN [Shi *et al.*, 2025]	8.54	5.48	53	46	43	7.62	6.31	40	36	34	8.02	6.65	37	32	30
Sim2Sim [Wu *et al.*, 2024a]	11.18	4.67	61	52	44	10.69	6.07	52	43	36	10.69	6.07	52	43	36
CM2 [Georgakis *et al.*, 2022]	12.05	6.10	51	43	35	11.54	7.02	42	34	28	13.90	7.70	39	31	24
WS-MGMAP [Chen *et al.*, 2022]	10.12	5.65	52	47	43	10.00	6.28	48	39	34	12.30	7.11	45	35	28
CWP-CMA [Hong *et al.*, 2022]	11.47	5.20	61	51	45	10.90	6.20	52	41	36	10.90	6.20	52	41	36
CWP-BERT [Hong *et al.*, 2022]	12.50	5.02	59	50	44	12.23	5.74	53	44	39	12.23	5.74	53	44	39
MV-Topo [Liu *et al.*, 2024]	11.25	3.69	71.3	65.5	59.7	11.85	4.72	63	57.5	49.1	12.71	4.94	60	54	48
ETPNav [An *et al.*, 2024]	11.78	3.95	72	66	59	11.99	4.71	65	57	49	11.99	4.71	65	57	49
FILD-Nav(Ours)	10.51	**3.88**	**76**	**67**	**61**	11.01	**4.64**	**72**	**59**	**52**	11.01	**4.64**	**72**	**59**	**52**

Table 1: Performance comparison on the VLN-CE dataset.

Module Configuration		**Val Seen**					**Val Unseen**				
Waypoint	**Node**	TL↓	NE↓	OSR↑	SR↑	SPL↑	TL↓	NE↓	OSR↑	SR↑	SPL↑
✓	✓	11.47	5.20	61	51	45	10.9	6.2	52	41	36
	✓	11.5	5.15	62	53	48	10.5	6.1	53	43	39
✓		11.78	3.95	72	66	59	11.99	4.71	65	57	49
✓	✓	10.51	3.80	**79**	**68**	**62**	11.01	4.60	**72**	**59**	**52**

Table 2: Ablation Study on Landmark Feature Integration for Navigation Performance.

Feature Input		**Waypoint Performance**			
RGBD	Landmark	$\|\Delta\|\downarrow$	%Open↑	$d_C\downarrow$	$d_H\downarrow$
✓		1.40	82.87	1.05	2.01
✓	✓	1.39	**84.05**	1.02	**1.81**

Table 3: Ablation Study on Waypoint Prediction with Instruction Landmark Features.

4.3 Ablation Study

On the unseen validation split (*val-unseen*), FILD-Nav exhibits clear improvements, with SR and SPL increasing by 2% and 3%, respectively, and OSR improving by 7%. These results demonstrate the robustness and generalization capability of our method in previously unseen environments, and validate the effectiveness of incorporating instruction-level landmark information.

In particular, compared with MV-Topo, which also integrates semantic information into visual topological maps, FILD-Nav achieves superior performance across all evaluation metrics. While MV-Topo directly extracts semantic cues from visual observations and embeds them into topological nodes, our method first derives landmark information from natural language instructions and then refines these landmarks using actual visual observations. By leveraging instruction-grounded landmarks to guide waypoint prediction and embedding them into topological nodes, our approach produces waypoint proposals that are more closely aligned with task semantics, resulting in more effective navigation behavior. These improvements indicate that instruction-grounded landmarks provide complementary semantic cues beyond visual-only representations.

Waypoint Prediction. Table 2 presents an ablation study on the R2R-CE dataset to evaluate the impact of instruction landmark features on waypoint prediction quality. We assess waypoint quality using four metrics: $|\Delta|$, the absolute difference between the numbers of target and predicted waypoints; $\%\mathrm{Open}$, the proportion of predicted waypoints located in obstacle-free open space; $\mathrm{d_C}$, the average distance error between the target and predicted waypoint sets; and $\mathrm{d_H}$, the maximum distance error between the two sets.

In the first setting, the waypoint predictor uses only RGB-D visual features as input. In the second setting, instruction landmark features are incorporated in addition to RGB-D features. While maintaining comparable performance in $|\Delta|$, the augmented model improves $\%\mathrm{Open}$ by 1.2%, indicating that waypoint obstruction and off-grid predictions are more effectively mitigated when landmark semantics are introduced. Moreover, both $\mathrm{d_C}$ and $\mathrm{d_H}$ decrease by 0.2 meters. This reduction is particularly important, as large waypoint deviations can severely degrade downstream navigation performance.

Overall, these results demonstrate that incorporating instruction landmark features enables the waypoint predictor to generate spatially more accurate and reliable waypoint proposals that are better aligned with task-relevant targets.

Navigation Performance with Landmark Feature Integration. To further investigate the effectiveness of instruction landmark features on end-to-end navigation, we conduct ablation studies on two key components: the waypoint prediction module and the node encoding module in the cross-modal planner. We evaluate individual and combined configurations on the R2R-CE dataset, reporting results on both the *val-seen* and *val-unseen* splits. The results are summarized in Table 3.

In the first setting, landmark features are excluded from

both waypoint prediction and node encoding. In the second setting, landmark features are introduced only in the waypoint prediction stage. Compared to the baseline, all evaluation metrics show consistent improvements, with trajectory length (TL) reduced by 0.9 meters and SPL increased by 3%. This indicates that incorporating landmark semantics into waypoint prediction helps generate more efficient navigation trajectories.

In the third setting, landmark features are added exclusively to the node encoding in the cross-modal planning module. While most metrics remain largely unchanged, the oracle success rate (OSR) improves by 7% and the success rate (SR) increases by 3%. This suggests that enriching node representations with instruction-grounded landmark semantics enhances alignment between topological nodes and task instructions, thereby facilitating more effective long-horizon planning.

Finally, when landmark features are incorporated into both waypoint prediction and node encoding, the agent achieves the best overall performance across all metrics. This configuration demonstrates improved instruction grounding and stronger generalization capability in complex and previously unseen environments.

5 Conclusion

This paper proposes FILD-Nav, a novel vision-and-language navigation framework that integrates instruction landmark features for continuous environments. By explicitly incorporating task-relevant landmark semantics into both the waypoint prediction module and the topological node representations, FILD-Nav improves the quality of waypoint generation and enhances cross-modal planning. As a result, the predicted waypoints and planned navigation paths are more closely aligned with the intended task goals. Extensive experimental results demonstrate the effectiveness of the proposed approach, with particularly notable improvements in previously unseen environments. Specifically, FILD-Nav achieves a 7% gain in Oracle Success Rate (OSR), a 2% increase in Success Rate (SR), and a 3% improvement in Success weighted by Path Length (SPL). These results indicate that FILD-Nav exhibits strong generalization capability and competitiveness when navigating novel and complex environments.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (Grant Nos. 62262066 and 62072071) and the Excellent Young Teachers' Research Startup Fund of XJNU(Grant No. XJNU202214).

References

[An *et al.*, 2024] Dong An, Hanqing Wang, Wenguan Wang, Zun Wang, Yan Huang, Keji He, and Liang Wang. Etpnav: Evolving topological planning for vision-language navigation in continuous environments. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.

[Anderson *et al.*, 2018] Peter Anderson, Qi Wu, Damien Teney, Jake Bruce, Mark Johnson, Niko Sünderhauf, Ian Reid, Stephen Gould, and Anton Van Den Hengel. Vision-and-language navigation: Interpreting visually-grounded navigation instructions in real environments. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3674–3683, 2018.

[Asadi *et al.*, 2019] Mahsa Asadi, Mohammad Sadegh Talebi, Hippolyte Bourel, and Odalric-Ambrym Maillard. Model-based reinforcement learning exploiting state-action equivalence. In *Asian Conference on Machine Learning*, pages 204–219. PMLR, 2019.

[Bao *et al.*, 2025] Xiaoan Bao, Zhiqiang Lv, and Biao Wu. Enhancing large language models with rag for visual language navigation in continuous environments. *Electronics*, 14(5):909, 2025.

[Cai *et al.*, 2024] Wenzhe Cai, Siyuan Huang, Guangran Cheng, Yuxing Long, Peng Gao, Changyin Sun, and Hao Dong. Bridging zero-shot object navigation and foundation models through pixel-guided navigation skill. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pages 5228–5234. IEEE, 2024.

[Chen *et al.*, 2021] Kevin Chen, Junshen K Chen, Jo Chuang, Marynel Vázquez, and Silvio Savarese. Topological planning with transformers for vision-and-language navigation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11276–11286, 2021.

[Chen *et al.*, 2022] Peihao Chen, Dongyu Ji, Kunyang Lin, Runhao Zeng, Thomas Li, Mingkui Tan, and Chuang Gan. Weakly-supervised multi-granularity map learning for vision-and-language navigation. *Advances in Neural Information Processing Systems*, 35:38149–38161, 2022.

[Chib and Singh, 2024] Pranav Singh Chib and Pravendra Singh. Enhancing trajectory prediction through self-supervised waypoint distortion prediction. In *International Conference on Machine Learning*, pages 8403–8416. PMLR, 2024.

[Dai *et al.*, 2024] Guangzhao Dai, Jian Zhao, Yuantao Chen, Yusen Qin, Hao Zhao, Guosen Xie, Yazhou Yao, Xiangbo Shu, and Xuelong Li. Unitedvln: Generalizable gaussian splatting for continuous vision-language navigation. *arXiv preprint arXiv:2411.16053*, 2024.

[Georgakis *et al.*, 2022] Georgios Georgakis, Karl Schmeckpeper, Karan Wanchoo, Soham Dan, Eleni Miltsakaki, Dan Roth, and Kostas Daniilidis. Cross-modal map learning for vision and language navigation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 15460–15470, 2022.

[He *et al.*, 2024] Keji He, Ya Jing, Yan Huang, Zhihe Lu, Dong An, and Liang Wang. Memory-adaptive vision-and-language navigation. *Pattern Recognition*, 153:110511, 2024.

[Hong *et al.*, 2022] Yicong Hong, Zun Wang, Qi Wu, and Stephen Gould. Bridging the gap between learning in discrete and continuous environments for vision-and-language navigation. In *Proceedings of the IEEE/CVF*

conference on computer vision and pattern recognition, pages 15439–15449, 2022.

[Huang *et al.*, 2022] Chenguang Huang, Oier Mees, Andy Zeng, and Wolfram Burgard. Visual language maps for robot navigation. *arXiv preprint arXiv:2210.05714*, 2022.

[Krantz and Lee, 2022] Jacob Krantz and Stefan Lee. Sim-2-sim transfer for vision-and-language navigation in continuous environments. In *European conference on computer vision*, pages 588–603. Springer, 2022.

[Krantz *et al.*, 2020] Jacob Krantz, Erik Wijmans, Arjun Majumdar, Dhruv Batra, and Stefan Lee. Beyond the nav-graph: Vision-and-language navigation in continuous environments. In *European Conference on Computer Vision*, pages 104–120. Springer, 2020.

[Krantz *et al.*, 2021] Jacob Krantz, Aaron Gokaslan, Dhruv Batra, Stefan Lee, and Oleksandr Maksymets. Waypoint models for instruction-guided navigation in continuous environments. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15162–15171, 2021.

[Li *et al.*, 2023] Xiangyang Li, Zihan Wang, Jiahao Yang, Yaowei Wang, and Shuqiang Jiang. Kerm: Knowledge enhanced reasoning for vision-and-language navigation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2583–2592, 2023.

[Lin *et al.*, 2024] Bingqian Lin, Yunshuang Nie, Ziming Wei, Yi Zhu, Hang Xu, Shikui Ma, Jianzhuang Liu, and Xiaodan Liang. Correctable landmark discovery via large models for vision-language navigation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(12):8534–8548, 2024.

[Liu *et al.*, 2024] Ruonan Liu, Ping Kong, and Weidong Zhang. Multiple visual features in topological map for vision-and-language navigation. In *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 7742–7749. IEEE, 2024.

[Qiao *et al.*, 2024] Yanyuan Qiao, Qianyi Liu, Jiajun Liu, Jing Liu, and Qi Wu. Llm as copilot for coarse-grained vision-and-language navigation. In *European Conference on Computer Vision*, pages 459–476. Springer, 2024.

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

[Raychaudhuri *et al.*, 2021] Sonia Raychaudhuri, Saim Wani, Shivansh Patel, Unnat Jain, and Angel Chang. Language-aligned waypoint (law) supervision for vision-and-language navigation in continuous environments. In *Proceedings of the 2021 conference on empirical methods in natural language processing*, pages 4018–4028, 2021.

[Savva *et al.*, 2019] Manolis Savva, Abhishek Kadian, Oleksandr Maksymets, Yili Zhao, Erik Wijmans, Bhavana Jain, Julian Straub, Jia Liu, Vladlen Koltun, Jitendra Malik, et al. Habitat: A platform for embodied ai research. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9339–9347, 2019.

[Shi *et al.*, 2025] Xiangyu Shi, Zerui Li, Wenqi Lyu, Jiatong Xia, Feras Dayoub, Yanyuan Qiao, and Qi Wu. Smartway: Enhanced waypoint prediction and backtracking for zero-shot vision-and-language navigation. *arXiv preprint arXiv:2503.10069*, 2025.

[Tan *et al.*, 2024] Sinan Tan, Kuankuan Sima, Dunzheng Wang, Mengmeng Ge, Di Guo, and Huaping Liu. Self-supervised 3-d semantic representation learning for vision-and-language navigation. *IEEE Transactions on Neural Networks and Learning Systems*, 36(4):6738–6751, 2024.

[Wang *et al.*, 2024a] Jiawei Wang, Teng Wang, Wenzhe Cai, Lele Xu, and Changyin Sun. Boosting efficient reinforcement learning for vision-and-language navigation with open-sourced llm. *IEEE Robotics and Automation Letters*, 2024.

[Wang *et al.*, 2024b] Jiawei Wang, Teng Wang, Lele Xu, Zichen He, and Changyin Sun. Discovering intrinsic subgoals for vision-and-language navigation via hierarchical reinforcement learning. *IEEE Transactions on Neural Networks and Learning Systems*, 36(4):6516–6528, 2024.

[Wen *et al.*, 2024] Shuhuan Wen, Simeng Gong, Ziyuan Zhang, F Richard Yu, and Zhiwen Wang. Vision-and-language navigation based on history-aware cross-modal feature fusion in indoor environment. *Knowledge-Based Systems*, 305:112610, 2024.

[Wijmans *et al.*, 2019] Erik Wijmans, Abhishek Kadian, Ari Morcos, Stefan Lee, Irfan Essa, Devi Parikh, Manolis Savva, and Dhruv Batra. Dd-ppo: Learning near-perfect pointgoal navigators from 2.5 billion frames. *arXiv preprint arXiv:1911.00357*, 2019.

[Wu *et al.*, 2024a] Siying Wu, Xueyang Fu, Feng Wu, and Zheng-Jun Zha. Vision-and-language navigation via latent semantic alignment learning. *IEEE Transactions on Multimedia*, 26:8406–8418, 2024.

[Wu *et al.*, 2024b] Wansen Wu, Tao Chang, Xinmeng Li, Quanjun Yin, and Yue Hu. Vision-language navigation: a survey and taxonomy. *Neural Computing and Applications*, 36(7):3291–3316, 2024.

[Xu *et al.*, 2025] Yunzhe Xu, Yiyuan Pan, Zhe Liu, and Hesheng Wang. Flame: Learning to navigate with multimodal llm in urban environments. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 9005–9013, 2025.

[Zheng *et al.*, 2025] Qi Zheng, Daqing Liu, Chaoyue Wang, Jing Zhang, Dadong Wang, and Dacheng Tao. Esceme: Vision-and-language navigation with episodic scene memory. *International Journal of Computer Vision*, 133(1):254–274, 2025.

[Zhou *et al.*, 2024] Gengze Zhou, Yicong Hong, and Qi Wu. Navgpt: Explicit reasoning in vision-and-language navigation with large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 7641–7649, 2024.