Think Hierarchically, Act Dynamically: Hierarchical Multi-modal Fusion and Reasoning for Vision-and-Language Navigation

Junrong Yue
City University of Hong Kong,
Dongguan Campus
China

Chuan Qin
The University of Melbourne
Australia

Bo Li
Tsinghua University & Baidu Inc.
China

Wenxin Zhang
University of Chinese Academy of
Science
China

Xinlei Yu
National University of Singapore
Singapore

Xiaomin Lie
City University of Hong Kong,
Dongguan Campus
China

Zhendong Zhao
University of Chinese Academy of
Science
China

Yifan Zhang*
City University of Hong Kong,
Dongguan Campus
China

Abstract

Vision-and-Language Navigation (VLN) aims to enable embodied agents to follow natural language instructions and reach target locations in real-world environments. While prior methods often rely on either global scene representations or object-level features, these approaches are insufficient for capturing the complex interactions across modalities required for accurate navigation. In this paper, we propose a Multi-level Fusion and Reasoning Architecture (MFRA) to enhance the agent's ability to reason over visual observations, language instructions and navigation history. Specifically, MFRA introduces a hierarchical fusion mechanism that aggregates multi-level features—ranging from low-level visual cues to high-level semantic concepts—across multiple modalities. We further design a reasoning module that leverages fused representations to infer navigation actions through instruction-guided attention and dynamic context integration. By selectively capturing and combining relevant visual, linguistic, and temporal signals, MFRA improves decision-making accuracy in complex navigation scenarios. Extensive experiments on benchmark VLN datasets including REVERIE, R2R, and SOON demonstrate that MFRA achieves superior performance compared to state-of-the-art methods, validating the effectiveness of multi-level modal fusion for embodied navigation.

CCS Concepts

• **Information systems** → **Information systems applications**;
• **Computing methodologies** → **Knowledge representation and reasoning**.

Keywords

Vision-and-Language Navigation; Hierarchical Multi-Modal Fusion; Instruction-Guided Attention; Dynamic Context Integration

1 Introduction

Vision-and-Language Navigation (VLN) [4, 19, 34, 35, 63, 66] represents a key research area in embodied Artificial Intelligence (AI),

where autonomous agents are tasked with processing natural language instructions, interpreting dynamic 3D surroundings and performing accurate navigation actions. This capability is essential for applications ranging from assistive robotics to augmented reality, as it bridges the gap between high-level linguistic commands and low-level spatial reasoning. Unlike conventional navigation systems [12, 13, 25, 43] that rely solely on visual inputs or predefined waypoints, VLN demands seamless integration of multi-modal inputs, including visual perception, linguistic comprehension, and temporal context, to achieve robust performance in real-world scenarios. The complexity of VLN lies in its requirement for agents to ground abstract language concepts into concrete visual and spatial representations, making it a challenging but critical area of study.

Early VLN approaches adopted polarized representation strategies, either relying exclusively on global scene embeddings [8, 17, 19, 20, 37] or object-centric features [1, 10, 33, 38, 64]. As Figure 1 demonstrates through a directional selection scenario where only candidate 3 aligns with the instruction, scene-level approaches encoded panoramic views as unitary descriptors via pretrained vision models, while object-based methods decomposed environments into discrete entity features. However, these approaches created two distinct failure modes: (1) holistic features introduced by the global scene representation lack discriminative power in semantically homogeneous environments, e.g., differentiating "bedroom" for candidates 1 and 3, while (2) object-level features overlook contextual relationships critical for navigation, e.g., "window" and "mirror" with similar object-level features, proving insufficient as both candidates share common features. Furthermore, both paradigms suffered from RNN-based history compression [11, 30, 46, 51, 54, 56], which may lead to the loss of the temporal information. Specifically, the fixed-size representations used by RNNs may cause additional information loss, particularly in long-horizon navigation tasks [58]. This bottleneck hindered the agent's ability to leverage rich temporal context, resulting in suboptimal navigation performance, especially in environments with complex layouts or ambiguous instructions.

Recognizing these limitations, recent advancements [5, 8, 10, 16, 18, 21, 28, 36, 37, 50, 57, 62] have attempted hybrid solutions by
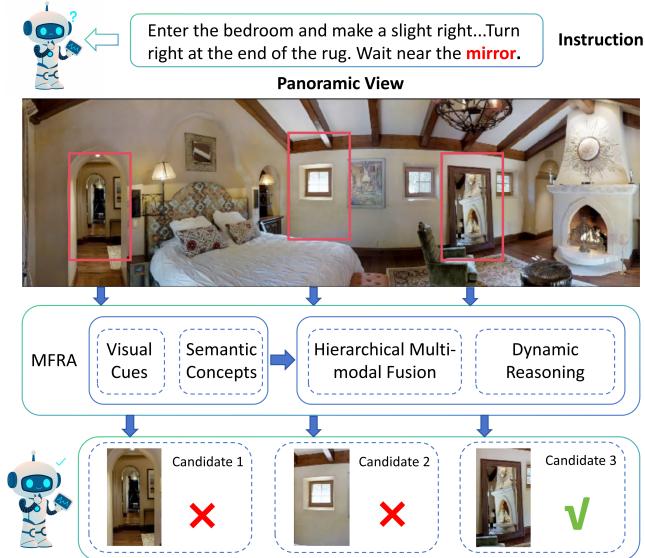
---

*Corresponding author.

Figure 1: Illustration of MFRA selected navigable candidates, which provides crucial information such as attributes and relationships between objects for further VLN reasoning module. Best viewed in color.

integrating global and object features. For instance, transformer-based methods [8, 10, 37] preserve variable-length histories and integrate topological mapping, e.g., DUET [10]), which improve the navigation performance through external knowledge but at increased computational cost. Advanced architectures [21, 50] better integrate language, observations, and history for multi-modal reasoning, but performance still degrades with longer instructions due to growing path complexity. However, the following issues still need to be addressed for the current methods. First, static feature weighting fails to adapt to varying instruction demands [5], e.g., prioritizing objects when hearing "Find the red mug" but scenes for "Enter the sunlit lounge". Second, treating all feature levels equally ignores the human-like reasoning hierarchy [18] where low-level cues (object shapes) inform mid-level concepts (furniture arrangements) that scaffold high-level semantics (room purposes). Third, most hybrid models still rely on RNNs or vanilla transformers for history encoding [16], causing performance drops in long-horizon SOON [63] tasks compared to short R2R [4] trajectories, which is inadequate for long-horizon tasks. Therefore, these hybrid approaches cannot effectively combine low-level sensory cues with high-level cognitive signals to mimic human-like reasoning.

To overcome these challenges, we propose the Multi-level Fusion and Reasoning Architecture (MFRA). Firstly, to address the inflexibility of static feature weighting mechanisms in adapting to diverse instructional requirements, MFRA introduces a hierarchical fusion mechanism that efficiently aggregates features across multiple levels of abstraction, from low-level visual cues to high-level semantic concepts. Secondly, to deal with the neglect of hierarchical cognitive processing that progressively integrates object-level attributes with spatial configurations and semantic contexts, we design a dynamic reasoning module that leverages instruction-guided

attention to align object features with linguistic context, which benefits action prediction in challenging scenarios. Third, to mitigate the limitations of conventional temporal encoding architectures in maintaining coherent historical representations for extended navigation sequences, MFRA employs contrastive learning, which enforces similarity between related historical states while distinguishing irrelevant ones by projecting multi-modal input into a shared embedding space, thus enabling seamless integration of visual, linguistic, and temporal modalities.

In this work, we introduce the MFRA, a novel framework designed to enhance embodied agents' reasoning capabilities by effectively integrating visual observations, language instructions, and navigation history. The architecture employs a hierarchical fusion mechanism that systematically combines multi-level features spanning from low-level visual cues to high-level semantic concepts across different modalities through **thinking like a human**. Building upon these fused representations, MFRA incorporates a sophisticated reasoning module that utilizes instruction-guided attention and dynamic context integration to infer optimal navigation actions to **act like a human**. Through its ability to selectively capture and combine the most relevant visual, linguistic, and temporal signals, MFRA achieves superior decision-making accuracy in complex navigation scenarios, as demonstrated through extensive experimental validation on standard benchmarks. In summary, we make the following contributions:

- We present a novel hierarchical fusion mechanism that systematically combines multi-level features to address the lack of discriminative power in global representations.
- We develop a dynamic reasoning module that enhances cross-modal alignment through instruction-guided attention to mitigate the limitations of object-centric approaches.
- We conduct extensive experiments to validate the effectiveness of our method and show that it outperforms existing methods with a better generalization ability.

2 Related Work

Vision-and-Language Navigation. Vision-and-Language Navigation (VLN) [4, 9, 10, 21, 38, 49, 50] has emerged as a pivotal research domain due to its transformative potential in applications ranging from healthcare robotics to personalized assistive systems. The core challenge requires agents to interpret diverse natural language instructions including step-by-step directives [26, 51], high-level goals [37, 63], and interactive dialogues [35] and translate them into precise navigation trajectories. Early methodologies [20, 51] relied on Recurrent Neural Networks (RNNs) to compress historical observations and actions into fixed-dimensional state vectors, a paradigm later enhanced through richer visual encoding. Notable advancements include scene-level and object-level relational modeling [21] and topological maps [6] for long-term spatial memory.

The advent of Vision-and-Language Pretraining (VLP) [41, 47, 59] catalyzed a paradigm shift toward Transformer-based architectures. Pioneering works like MiniVLN [65] introduced a two-stage distillation model covering both the pre-training and fine-tuning phases, while AirBERT [19] scaled up multi-modal training data to strengthen modality interactions and RecBERT [27] utilizes the [CLS] token within the transformer as a recurrent state to record

navigation history. Subsequent innovations refined cross-modal fusion: VLNBERT [22] incorporated recurrent temporal modeling, HAMT [8] unified language-history-observation sequences via cross-modal attention, and DUET [10] pioneered graph transformers to bridge local-global spatial reasoning. In this work, we leverage a hierarchical fusion mechanism that aggregates multi-level features across multiple modalities to improve the generalization ability and facilitate the alignment between vision and language for VLN.

Vision-and-Language Models with Fused Representations. The release of the Room-to-Room (R2R) benchmark [4] marked a pivotal advancement in VLN research, spurring the development of numerous computational models tailored for discrete environments. Initial approaches, exemplified by the Seq2Seq architecture [4] and the RCM framework [6], primarily employed imitation learning and reinforcement learning paradigms within standard egocentric visual settings. Subsequent methodological innovations, including CLIP-ViL [44], integrated enhanced visual representations derived from large-scale pretrained models such as CLIP [41, 47].

A significant research trajectory subsequently emerged around the efficient encoding of temporal context, with architectures like VLN-BERT [22] adopting recurrent transformer mechanisms. More recent work [2, 8] has investigated the incorporation of structured spatial representations, including topological and metric maps, to augment navigational context. Concurrent with these architectural developments, efforts such as ScaleVLN [53] have sought to address data scarcity through large-scale training corpus expansion.

Most recently, the field has witnessed a paradigm shift toward leveraging Large Language Models (LLMs) for VLN, as evidenced by [7, 29, 32, 61]. In this work, we design a reasoning module that leverages hierarchical fused representations to infer navigation actions through instruction-guided attention and dynamic context integration to benefit VLN tasks.

3 Methodology

In this work, we propose a novel and comprehensive framework, termed **Multi-level Fusion and Reasoning Architecture (MFRA)**, to address the challenges of multi-modal understanding in VLN tasks. VLN requires embodied agents to perceive their visual surroundings, comprehend natural language instructions, and reason over their navigation history to make informed decisions in complex environments. However, effectively integrating these heterogeneous modalities—particularly under ambiguous semantics and dynamic scenes—remains a non-trivial problem. To this end, MFRA is designed to perform hierarchical fusion and semantic reasoning across visual, linguistic, and temporal dimensions. By constructing a unified representation space through multi-level cross-modal interactions, the proposed architecture enables the agent to better align instruction semantics with visual observations and contextual history. This section formally defines the VLN problem setting and elaborates on the key components and mechanisms of MFRA in detail.

Problem Formulation. The VLN task is formally modeled as a topological graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where each node $v \in \mathcal{V}$ denotes a navigable viewpoint and each edge $e \in \mathcal{E}$ encodes the connectivity between adjacent viewpoints. The navigation episode begins at a source node v_s, and the agent is instructed to reach a target node v_g by following a natural language instruction $I = \{w_1, w_2, ..., w_T\}$, where w_t represents the t-th token in the instruction sequence. At each time step t, the agent receives a panoramic observation $S_t = \{s_i^t\}_{i=1}^{36}$, consisting of 36 single-view images evenly distributed across the viewing sphere. Each view s_i^t is represented by a visual feature vector $v_i^t \in \mathbb{R}^{d_v}$ along with an associated orientation embedding $o_i^t \in \mathbb{R}^{d_o}$, capturing the spatial direction of the view. For instruction settings involving object grounding, an additional set of object-level features is provided as $O_t = \{o_j^t\}_{j=1}^{N}$, where each o_j^t corresponds to a detected or annotated object within the current panorama. The objective is to learn a navigation policy $\pi(a_t \mid S_t, I, H_t)$ that predicts an action $a_t \in \mathcal{A}$ at each timestep, conditioned on the current observation S_t, the instruction I, and the accumulated navigation history H_t. The agent continues to interact with the environment until it selects a termination action or reaches a predefined maximum number of steps.

3.1 Overview of MFRA

To address the challenges of multi-modal semantic alignment and context-aware reasoning in VLN, we propose the **Multi-level Fusion and Reasoning Architecture (MFRA)**. MFRA is designed to jointly model panoramic visual observations, natural language instructions, and the agent's navigation history through a hierarchical fusion pipeline and a dynamic reasoning mechanism.

The overall architecture of MFRA consists of three components: (1) *Multi-level Feature Extraction*, which encodes raw visual inputs, instruction tokens, and trajectory history into structured and semantically aligned representations; (2) *Hierarchical Multi-modal Fusion*, which leverages a DIRformer-based structure to align and integrate information across modalities and semantic levels; and (3) *Dynamic Reasoning Module*, which performs instruction-guided attention and context-aware decision making for action prediction.

In the first stage, we utilize a pretrained multi-modal model, CLIP [40], to obtain aligned visual-language features for both panoramic views and instruction tokens. These multi-level features are then fed into a multi-stage Transformer backbone based on the Dynamic IR Transformer (DIRformer), which enables fine-grained visual-language interaction across different abstraction levels. The fused representation is passed to a reasoning module that adaptively integrates temporal context and instruction relevance to infer the most appropriate navigation action.

3.2 Multi-modal Feature Extraction

To enable effective fusion and reasoning, we extract semantically aligned representations from three modalities: vision, language, and navigation history.

At each timestep t, the agent receives a **panoramic visual observation** composed of 36 discrete single-view images $S_t = \{s_i^t\}_{i=1}^{36}$. Each view s_i^t is encoded using the vision encoder of CLIP [40], producing a set of visual feature vectors $V_t = \{v_i^t\}_{i=1}^{36}$, where $v_i^t \in \mathbb{R}^{d_v}$, which are embedded in a shared vision-language space.
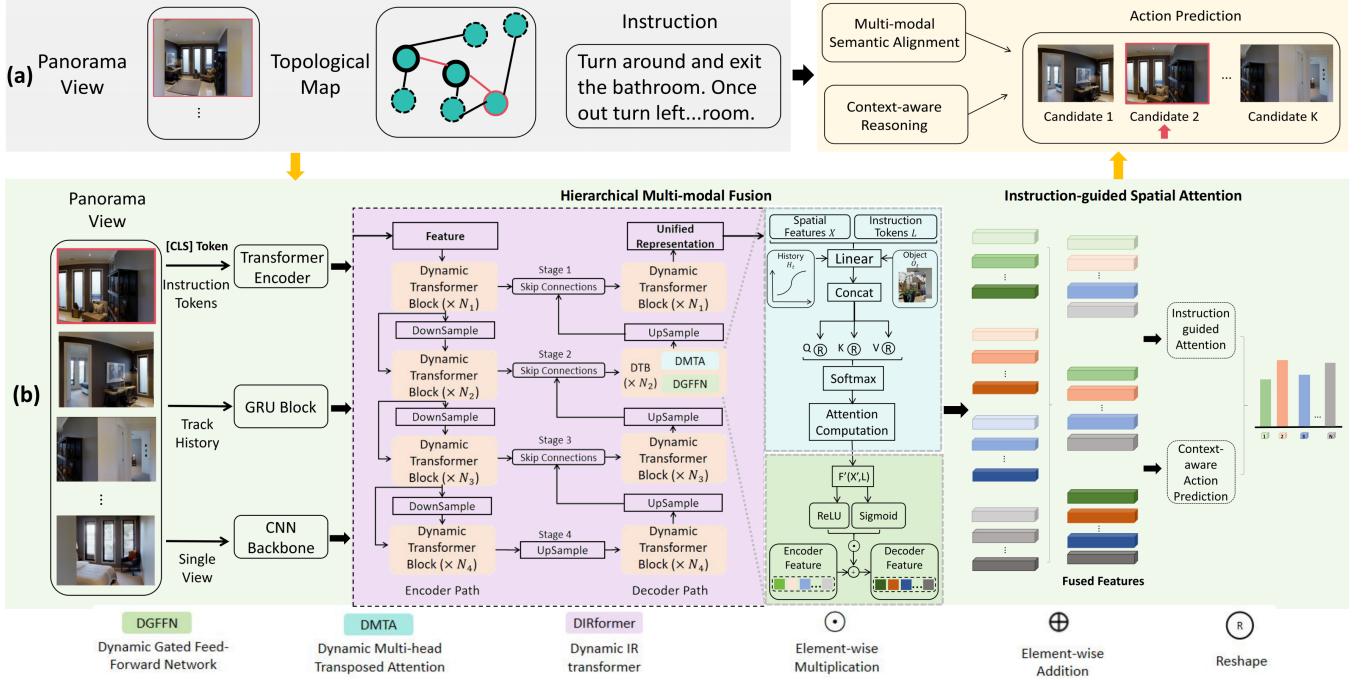
Figure 2: The overall pipeline. (a) The baseline method uses a dual-scale graph transformer to encode the panoramic view, the topological map, and the instruction for action prediction. (b) Our approach incorporates multi-level feature fusion as model input. The representations of each candidate view are obtained with the hierarchical multi-modal fusion, the instruction-guided spatial attention module and the context-aware interaction module. Best viewed in color.

The **natural language instruction**, represented as a token sequence $I = \{w_1, w_2, ..., w_T\}$, is embedded using the CLIP text encoder. This yields token-level embeddings $L = \{l_j\}_{j=1}^{T}$, along with a sentence-level representation $l_{\text{cls}} \in \mathbb{R}^{d_l}$ that captures global semantic intent. The visual and linguistic embeddings produced by CLIP are inherently aligned in a joint embedding space.

For scenarios involving object-level grounding, such as in the REVERIE and SOON datasets, we further extract **object-centric region** features $O_t = \{o_j^t\}_{j=1}^{N}$ from each panoramic observation. These features are derived from region proposals generated by a pretrained object detector, e.g., Faster R-CNN [42], and are projected into the same embedding space via a learnable adapter to maintain representational consistency with CLIP.

To capture the temporal evolution of the navigation process, we maintain a recurrent history embedding H_t, which summarizes the trajectory observed up to time t. At each step, the current visual observation and its instruction-attended context are fused through a learnable function $\phi(V_t, L)$, and the result is passed into a Gated Recurrent Unit (GRU) [14] to update the temporal state:

$$h_t = \text{GRU}(h_{t-1}, \phi(V_t, L)).$$

The resulting representations V_t, L, O_t, and H_t serve as the foundational multi-modal features, which are then fed into the hierarchical fusion module for joint representation learning and downstream navigation decision-making.

Language Encoding. We encode the instruction I using a Transformer encoder to obtain token-level features:

$$L = \text{Transformer}_{\text{Lang}}(I) = \{l_1, l_2, ..., l_T\}, \quad l_i \in \mathbb{R}^{d}$$

The first token's embedding l_{cls} is used as the global instruction context.

Visual Encoding. Each single view s_i^t is encoded via a CNN backbone and concatenated with its orientation:

$$v_i^t = \text{CNN}(s_i^t), \quad x_i^t = [v_i^t; o_i^t]$$

We denote $X_t = \{x_i^t\}_{i=1}^{36}$ as the full panoramic feature set.

History Encoding. The trajectory history is modeled using a GRU:

$$h_t = \text{GRU}(x_{1:t}), \quad H_t = \{h_i\}_{i=1}^{t}$$

3.3 Hierarchical Multi-modal Fusion

To resolve the inadequacy of uniform multi-modal feature integration, we emulate the human-like cognitive hierarchy, where low-level sensory cues progressively inform higher-order abstractions through designing a hierarchical fusion strategy inspired by DIRformer [23] across three semantic tiers: low-level (object shapes), mid-level (spatial arrangements), and high-level (semantic contexts). It enables fine-grained cross-modal alignment at multiple semantic levels and ensures each tier scaffolds the next through guided interactions.

Given the multi-modal features extracted in the previous stage, namely the visual embeddings $V_t = \{v_i^t\}$, instruction token embeddings $L = \{l_j\}$, object-level features $O_t = \{o_j^t\}$, and the temporal history representation H_t, we construct a multi-stage encoder-decoder architecture with dynamic interaction between modalities at each level. The fusion module is composed of four hierarchical stages. In the encoder path, each stage consists of a series of dynamic Transformer blocks, where each block includes a *Dynamic Multi-head Transposed Attention (DMTA)* layer followed by a *Dynamic Gated Feed-Forward Network (DGFFN)*. At stage s, the feature map $Z^{(s)}$ is updated as:

$$Z^{(s)} = \text{DGFFN}\left(\text{DMTA}(Z^{(s-1)}, L)\right),$$

where the DMTA module performs attention between spatial features and instruction tokens, and DGFFN controls the flow of information through dynamic gating [31, 45].

DMTA Module. The DMTA module projects the input spatial features X and instruction tokens L into query, key, and value spaces via linear transformations to leverage low-level features:

$$Q = W_q X, \quad K = W_k L, \quad V = W_v L,$$

and computes attention in a transposed manner:

$$A = \text{softmax}\left(\frac{QK^\top}{\sqrt{d}}\right), \quad \tilde{X} = AV.$$

The attended features $\tilde{X}$ are added to the input via residual connection, allowing instruction-guided semantic filtering at each scale.

DGFFN Module. The DGFFN module then applies non-linear transformation and dynamic gating to leverage mid-level features to aggregate object interactions (e.g., "sofa adjacent to table") into cohesive arrangements:

$$F_1 = \text{ReLU}(W_1 X'), \quad F_2 = \sigma(W_2 X'), \quad \text{DGFFN}(X') = F_1 \odot F_2,$$

where $\odot$ denotes element-wise multiplication. This allows selective activation of spatial-semantic patterns relevant to the current instruction and visual context.

Following each encoder stage, spatial resolution is reduced by downsampling (e.g. strided convolution), while channel dimensions are expanded to enable broader semantic abstraction. In the decoder path, a symmetric set of upsampling blocks reconstructs higher-resolution representations, with skip connections bridging encoder and decoder stages. The decoder feature at stage s is computed as:

$$\hat{Z}^{(s)} = \text{DGFFN}\left(\text{DMTA}(\text{Up}(\hat{Z}^{(s+1)}), L) \oplus Z^{(s)}\right),$$

where $\oplus$ denotes channel-wise concatenation. This structure ensures that both fine-grained and global cues are preserved throughout the decoding process. High-level reasoning is also performed by mid-level layouts and low-level features.

To further enhance high-level grounding and temporal reasoning, we incorporate object-level features O_t and trajectory history H_t into each fusion stage. These features are projected into token embeddings and fused via auxiliary DMTA layers:

$$\tilde{X}_o = \text{DMTA}(X, O_t), \quad \tilde{X}_h = \text{DMTA}(X, H_t),$$

with the combined output:

$$X'' = X' + \tilde{X}_o + \tilde{X}_h.$$

This enables the model to reason over fine-grained entities and sequential patterns jointly with visual and linguistic cues. In this way, the tiers achieve a bidirectional interaction: low-level details constrain mid-level arrangements, which in turn contextualize high-level goals. Conversely, high-level semantics (e.g., "find a bedroom") prune irrelevant mid-level configurations, which refine low-level feature extraction. This mimics human top-down and bottom-up reasoning, yielding a unified representation where all tiers co-evolve to resolve ambiguities and align with task objectives.

Algorithm 1: Pseudo-code of MFRA at Time Step t

Input: Panoramic visual observations $S_t = \{s_i^t\}_{i=1}^{36}$,
    instruction tokens $I = \{w_1, ..., w_T\}$, optional object
    features O_t, previous trajectory history H_{t-1}

Output: Predicted navigation action a_t

1. Multi-modal Feature Extraction:
    Encode S_t and I into visual features $V_t = \{v_i^t\}$ using
    CLIP visual encoder
    Encode I into token embeddings $L = \{l_j\}$ and global
    context l_{cls} using CLIP text encoder
    If object-level features exist, extract $O_t = \{o_j^t\}$ and
    project into shared space
    Fuse V_t and L into context vector $\phi(V_t, L)$
    Update temporal history embedding
    $h_t = \text{GRU}(H_{t-1}, \phi(V_t, L))$

2. Hierarchical Multi-modal Fusion (DIRformer):
    Initialize fused representation $Z^{(0)} = V_t$
    for *each encoder stage $s = 1$ to S* **do**
    | $Z^{(s)} \leftarrow \text{DGFFN}(\text{DMTA}(Z^{(s-1)}, L))$
    end
    Integrate object and history context:
    $\tilde{X}_o = \text{DMTA}(Z^{(S)}, O_t), \quad \tilde{X}_h = \text{DMTA}(Z^{(S)}, H_t)$
    $Z_{\text{fused}} = Z^{(S)} + \tilde{X}_o + \tilde{X}_h$

3. Dynamic Reasoning and Action Prediction:
    Compute instruction-guided spatial attention:
    $\bar{z}_t = \text{Attention}(Z_{\text{fused}}, l_{\text{cls}})$
    Form decision vector: $z_t^{\text{final}} = \text{FFN}([\bar{z}_t; h_t])$
    Predict action:
    $a_t = \arg\max_{a_k \in \mathcal{A}_t} \text{sim}(z_t^{\text{final}}, v_k)$

return a_t

3.4 Dynamic Reasoning Module

After hierarchical multi-modal fusion, the agent obtains a unified representation that encodes spatial visual information, instruction semantics, object cues, and temporal trajectory context. To translate this rich representation into actionable decisions, we design a **Dynamic Reasoning Module** that performs instruction-guided attention and context-aware action prediction.

At each navigation step t, the fused representation from the decoder output of the DIRformer module is denoted as $Z_t \in \mathbb{R}^{H \times W \times C}$, where each token corresponds to a spatial location in the agent's

panoramic view. To allow the agent to focus on regions that are semantically relevant to the instruction, we introduce an *Instruction-guided Spatial Attention* mechanism. Given the global instruction embedding $l_{\text{cls}} \in \mathbb{R}^C$, extracted from the [CLS] token of the CLIP text encoder, we compute attention weights over spatial features as:

$$\alpha_i = \frac{(W_r z_i)^\top (W_l l_{\text{cls}})}{\sqrt{d}}, \quad \eta = \text{softmax}(\{\alpha_i\}_{i=1}^{HW}), \quad \bar{z}_t = \sum_{i=1}^{HW} \eta_i z_i,$$

where z_i denotes the i-th token in Z_t, and W_r, W_l are learnable projection matrices. The resulting vector $\bar{z}_t \in \mathbb{R}^C$ captures the instruction-conditioned visual context at step t.

To incorporate temporal coherence and navigation progress, we further integrate the recurrent history embedding $h_t \in \mathbb{R}^C$, which summarizes all previous observations and actions up to timestep t. The final decision embedding is formed by concatenating the current attended context and historical representation:

$$z_t^{\text{final}} = \text{FFN}([\bar{z}_t; h_t]),$$

where FFN denotes a feed-forward network that projects the concatenated vector into a fixed-dimensional decision space.

The action prediction head is a classifier over the candidate navigable directions $\mathcal{A}_t$, including the stop action. Given the set of candidate view embeddings $\{v_k\} \subseteq V_t$, we compute their similarity to the decision embedding and apply softmax normalization:

$$\hat{a}_t = \arg\max_{a_k \in \mathcal{A}_t} \text{softmax}\left(\frac{(W_d z_t^{\text{final}})^\top v_k}{\sqrt{d}}\right),$$

where W_d is a learnable projection and each v_k corresponds to the CLIP-encoded feature of a candidate direction. The stop action is modeled similarly, using a dedicated stop embedding vector.

This dynamic reasoning process enables the agent to adaptively attend to instruction-relevant regions, leverage historical context, and evaluate the most appropriate next action in a unified decision space. The module is trained via behavior cloning, minimizing the negative log-likelihood of the expert action at each step.

3.5 Training Objective

The proposed MFRA framework is trained in a two-stage manner to fully exploit its multi-modal representation capacity and reasoning capability. The training objective comprises a primary navigation loss and several auxiliary objectives designed to improve semantic grounding, visual-language alignment, and reasoning robustness.

Supervised Action Learning. The core objective is to learn a navigation policy that maps the current observation and instruction to the optimal next action. We adopt behavior cloning to supervise the policy network using expert trajectories from the training dataset. Given the ground-truth action sequence $\{a_t^*\}_{t=1}^T$, the navigation loss is defined as the cross-entropy between the predicted action distribution and the ground-truth action:

$$\mathcal{L}_{\text{nav}} = -\sum_{t=1}^T \log \pi(a_t^* \mid S_t, I, H_t),$$

where $\pi(\cdot)$ denotes the learned policy, conditioned on the panoramic observation S_t, the instruction I, and the history embedding H_t.

Masked Language Modeling (MLM) [15]. To enhance instruction understanding and facilitate token-level attention, we apply a masked language modeling objective during pretraining. A random subset of instruction tokens is masked and the model is trained to reconstruct them using the fused visual context:

$$\mathcal{L}_{\text{MLM}} = -\sum_{j \in \mathcal{M}} \log P(w_j \mid S_t, I_{\backslash j}),$$

where $\mathcal{M}$ denotes the set of masked positions, and $I_{\backslash j}$ is the instruction with the j-th token masked.

Masked View Classification (MVC) [55]. To encourage semantic discrimination of visual tokens, we introduce a masked view classification task. A fraction of panoramic views are masked during training, and the model is required to predict their semantic categories:

$$\mathcal{L}_{\text{MVC}} = -\sum_{i \in \mathcal{V}_m} \log P(c_i \mid v_i),$$

where $\mathcal{V}_m$ is the set of masked view indices and c_i is the target semantic label of view v_i, obtained from an external visual classifier.

Object Grounding (OG). For tasks involving object-level grounding, such as REVERIE and SOON, we add an auxiliary loss that encourages the model to locate the correct object referred to by the instruction. The model is trained to predict the correct object o^* at the final location P_D:

$$\mathcal{L}_{\text{OG}} = -\log P(o^* \mid I, P_D),$$

Overall Loss. The total training objective is a weighted combination of the above components:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{nav}} + \lambda_1 \mathcal{L}_{\text{MLM}} + \lambda_2 \mathcal{L}_{\text{MVC}} + \lambda_3 \mathcal{L}_{\text{OG}},$$

where $\lambda_1, \lambda_2, \lambda_3$ are hyperparameters that balance the contributions of each auxiliary task. In our experiments, we empirically set $\lambda_1 = 1.0$, $\lambda_2 = 0.5$, and $\lambda_3 = 1.0$.

This joint training objective enables MFRA to learn semantically rich and well-aligned representations across modalities, while simultaneously grounding object references and enhancing instruction-conditioned reasoning.

4 Experiments

4.1 Datasets and Evaluation Metrics

We evaluate the proposed MFRA model on three widely-used VLN benchmarks: REVERIE [35], SOON [63], and R2R [4]. These datasets cover a wide range of navigation and grounding challenges with varying instruction complexity and visual ambiguity.

REVERIE provides high-level navigation instructions with an average length of 21 words. Each panoramic node includes predefined object bounding boxes, and the agent is required to identify the target object at the end of the navigation path. Trajectory lengths range from 4 to 7 steps. **SOON** features longer and more complex instructions (average length of 47 words) and does not include predefined bounding boxes. The agent must predict the object center location using object detectors [24]. Path lengths range from 2 to 21 steps. **R2R** consists of step-by-step navigation instructions with an average length of 32 words and an average trajectory length of

Table 1: Performance comparison with SOTA methods on the R2R dataset. All metrics are reported in %. TL is in meters.

Method	Val Seen						Val Unseen						Test Unseen					
	TL↓	OSR↑	SR↑	SPL↑	RGS↑	RGSPL↑	TL↓	OSR↑	SR↑	SPL↑	RGS↑	RGSPL↑	TL↓	OSR↑	SR↑	SPL↑	RGS↑	RGSPL↑
Seq2Seq	12.88	35.70	29.59	24.01	18.97	14.96	11.07	8.07	4.20	2.84	2.16	1.63	10.89	6.88	3.99	3.09	2.00	1.58
VLNBERT	13.44	53.90	51.79	47.96	38.23	35.61	16.78	35.02	30.67	24.90	18.77	15.27	15.68	32.91	29.61	23.99	16.50	13.51
AirBERT	15.16	49.98	47.01	42.34	32.75	30.01	18.71	34.51	27.89	21.88	18.23	14.18	17.91	34.20	30.28	23.61	16.83	13.28
HOP	13.80	54.88	53.76	47.19	38.65	33.85	16.46	36.24	31.78	26.11	18.85	15.73	16.38	33.06	30.17	24.34	17.69	14.34
DUET	13.86	73.68	71.75	63.94	57.41	51.14	22.11	51.07	46.98	33.73	32.15	23.03	21.30	56.91	52.51	36.06	31.88	22.06
NaviLLM	14.10	75.00	73.12	66.08	59.20	53.17	22.75	53.91	48.53	34.76	33.42	23.88	20.90	57.10	51.80	37.15	30.91	21.84
VLN-PETL	13.92	70.26	68.87	61.55	54.30	49.78	21.64	49.70	45.32	32.44	30.99	22.07	20.58	53.68	49.63	35.27	30.12	21.10
LaNA	14.08	71.92	70.04	62.89	56.01	50.10	21.88	50.13	45.76	33.55	31.67	22.33	21.02	55.11	50.47	36.12	31.08	21.49
ETPNav	13.71	72.84	71.01	63.00	56.39	51.25	21.93	51.32	46.21	33.80	31.21	22.17	21.10	55.43	51.06	36.04	31.26	21.73
MFRA (Ours)	**12.84**	**79.20**	**76.88**	**70.45**	**61.00**	**56.07**	**21.85**	**55.21**	**50.44**	**35.38**	**34.51**	**24.45**	**17.32**	**57.58**	**52.43**	**39.21**	**32.39**	**23.64**

Figure 3: Visualization of navigation examples. The sentence within the yellow box is the navigation instruction for the agent. We show a comparison where our MFRA chooses the right location while the baseline model makes the wrong choice. Best viewed in color.

6 steps. Unlike REVERIE and SOON, R2R focuses solely on spatial navigation without explicit object grounding.

We adopt standard VLN evaluation metrics, including: (1) **Trajectory Length (TL)**: average length of navigation paths; (2) **Navigation Error (NE)**: average distance (in meters) between the agent's final location and the goal; (3) **Success Rate (SR)**: percentage of episodes where NE is less than 3 meters; (4) **Oracle Success Rate (OSR)**: SR assuming an oracle stop policy; (5) **SPL**: SR weighted by path efficiency. For datasets with grounding tasks (REVERIE, SOON), we further report: (6) **Remote Grounding Success (RGS)**: proportion of successfully grounded instructions; (7) **RGSPL**: RGS weighted by path length. For all metrics except TL and NE, higher values indicate better performance.

We adopt CLIP-ViT-B/16 to extract aligned visual-language features for both panoramic views and instruction tokens. Object features on REVERIE are obtained using ViT-B/16 pretrained on ImageNet, while object bounding boxes on SOON are extracted using the BUTD detector [24]. The fusion module is implemented using a 4-stage DIRformer encoder-decoder architecture with DMTA and DGFFN blocks. Cross-modal interaction layers are initialized from LXMERT [48].

4.2 Experimental Results and Analysis

We compare our proposed MFRA model with a range of representative state-of-the-art methods on the R2R dataset, covering both traditional sequence-based baselines [4] and recent transformer-based or knowledge-enhanced architectures [10, 19, 22, 37], as well as newly proposed models such as NaviLLM [60], VLN-PETL [39], LaNA [52], and ETPNav [3]. The quantitative results on the validation seen, validation unseen, and test unseen splits are summarized in Table 1. The comparison results on other datasets can be found in the supplementary material.

The results presented in Table 1 comprehensively demonstrate the superiority of our proposed MFRA framework across all evaluation metrics on the REVERIE dataset. Compared with prior state-of-the-art methods, MFRA achieves consistent and significant improvements in both navigation accuracy and object grounding precision. Specifically, MFRA outperforms DUET by +5.13% in Success Rate (SR) and +6.51% in SPL on the validation seen split, while also achieving higher Remote Grounding Success (RGS) and RGSPL, confirming the effectiveness of our hierarchical fusion and semantic reasoning design. Moreover, compared with recent advances such as NaviLLM and ETPNav—which incorporate large language model priors and evolving topological planning—MFRA still shows superior performance, with a clear advantage of 2–5% on both SR and RGS across most splits. While VLN-PETL demonstrates efficiency through parameter-efficient tuning and LaNA introduces bi-directional language understanding, their navigation accuracy remains lower than that of MFRA, particularly under domain shifts and object ambiguity.

Notably, the performance drop from seen to unseen environments is much smaller than that of all baseline methods, indicating stronger generalization capabilities. This robustness stems from the use of CLIP-based multi-modal features trained with large-scale language supervision, and from our unified DIRformer fusion backbone which supports modality-invariant representation learning. Furthermore, the integration of dynamic multi-head transposed attention and gated feed-forward networks enables our model to selectively attend to instruction-relevant regions, improving cross-modal semantic alignment and decision confidence. Compared to object-centric approaches such as AirBERT or VLNBERT, MFRA benefits from a fact-level grounding mechanism that captures richer contextual information beyond static object labels. This allows the agent to reason over both fine-grained visual cues and high-level semantic references within instructions. Overall, the strong empirical results validate that MFRA is capable of learning robust, generalizable, and semantically grounded navigation policies, outperforming

existing methods under varying conditions of instruction complexity, scene diversity, and task ambiguity.

4.3 Ablation Study

To better understand the contribution of each component in our proposed MFRA framework, we conduct a comprehensive ablation study on the REVERIE validation unseen split. Specifically, we progressively remove or modify critical modules in the architecture and evaluate their impact on both navigation and grounding performance. The results are summarized in Table 2.

As shown in Table 2, removing the hierarchical DIRformer fusion module leads to the most significant performance degradation, with a 4.42% drop in SR and over 2.3% drop in RGSPL, confirming the central role of multi-level cross-modal reasoning in MFRA. Disabling the instruction-guided attention module also reduces SR and RGS, indicating that semantic alignment between instruction-level abstraction and visual context is critical for accurate decision-making. The exclusion of the history-fact interaction module leads to a moderate decline, suggesting that temporal information contributes meaningfully to global context reasoning. Finally, replacing CLIP-based fact features with conventional CNN+LSTM-based representations causes the largest overall performance drop in both navigation and grounding accuracy, demonstrating that pretrained multi-modal embeddings offer superior semantic alignment capabilities. These ablation results collectively verify that each module in MFRA—especially DIRformer fusion and CLIP-based representation—plays a vital role in enabling accurate and generalizable instruction-grounded navigation.

Table 2: Ablation study results on the R2R val unseen split.

Model Variant	SR↑	SPL↑	OSR↑	RGS↑	RGSPL↑
Full MFRA (Ours)	**50.44**	**35.38**	**55.21**	**34.51**	**24.45**
w/o DIRformer Fusion	46.02	32.01	51.87	31.08	22.13
w/o Instruction-guided Attention	47.12	33.10	52.44	32.28	22.86
w/o History-Fact Interaction	48.30	34.12	53.53	33.63	23.58
w/o CLIP-based Fact Representation	44.91	31.04	50.11	30.25	21.14

Table 3: Analysis of multi-modal feature contributions on the R2R val unseen split.

Feature Configuration	SR↑	SPL↑	OSR↑	RGS↑	RGSPL↑
All Modalities (Full MFRA)	**50.44**	**35.38**	**55.21**	**34.51**	**24.45**
w/o Visual Feature	41.27	27.65	47.04	28.12	19.90
w/o Language Feature	36.08	23.54	43.19	25.89	17.43
w/o History Feature	46.11	32.79	52.44	32.15	22.04
Visual + Language only	48.05	34.01	54.12	33.22	23.27
Visual + History only	43.37	30.12	50.86	30.84	21.06
Language + History only	40.90	28.23	48.21	29.14	20.36

4.4 Multi-modal Feature Contribution Analysis

To further investigate the contribution of individual modalities in MFRA, we conduct a series of controlled experiments by selectively removing vision, language, or history components from the fused representation. Specifically, we analyze how navigation performance is affected when one or more modalities are excluded from

the multi-modal reasoning process. The results on the REVERIE validation unseen split are reported in Table 3.

The results demonstrate that all three modalities—vision, language, and history—play complementary roles in supporting effective navigation and grounding. When the visual modality is removed, performance drops sharply across all metrics (-9.17% SR and -4.55% RGSPL), indicating that grounded visual perception is essential for scene understanding. Excluding the instruction input leads to the largest performance degradation (-14.36% SR and -7.02%), as language provides high-level semantic guidance that conditions the agent's decisions. The history feature also contributes meaningfully, particularly in grounding-oriented metrics, as removing it causes a -2.41% drop in RGSPL. We further analyze partial modality combinations. While visual and language features alone can support reasonable performance (48.05% SR), integrating history yields additional gains, highlighting the importance of temporal context. In contrast, using only language and history without visual grounding leads to significant performance degradation.

4.5 Cross-Dataset Evaluation on REVERIE and SOON

To further evaluate the generalization performance of MFRA, we conduct experiments on the REVERIE and SOON datasets, both of which pose greater challenges in object grounding and long-horizon instruction following. These datasets serve as rigorous benchmarks for testing multi-modal fusion, semantic reasoning, and policy robustness in unseen environments. Figure 4 illustrates the Success Rate (SR) and Remote Grounding Success (RGS) of MFRA in comparison with a broad set of representative baselines, including both classical transformer-based models such as VLNBERT, AirBERT, and DUET, as well as more recent approaches like NaviLLM, VLN-PETL, LaNA, and ETPNav. Across both datasets, MFRA consistently achieves the highest performance, demonstrating stronger instruction-following accuracy and grounding precision under complex and diverse scene conditions.
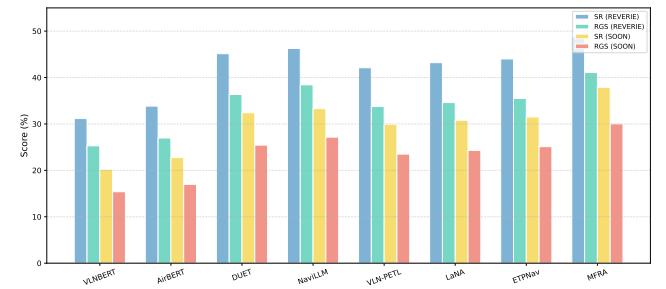

Figure 4: Performance comparison on REVERIE and SOON validation unseen splits. MFRA achieves consistent improvements in Success Rate (SR) and Remote Grounding Success (RGS) over baseline methods.

Notably, while methods such as NaviLLM and ETPNav benefit from large language model priors or topological planning, their performance still lags behind MFRA, particularly in generalizing to previously unseen environments. VLN-PETL and LaNA introduce

parameter-efficient adaptation and generative capabilities, but face limitations when dealing with fine-grained object semantics and long instruction spans. In contrast, MFRA leverages unified visual-language representations derived from CLIP, combined with a hierarchical DIRformer-based fusion mechanism, enabling effective cross-modal alignment and robust reasoning over both low-level visual cues and high-level semantic structures.

These results collectively demonstrate the superior generalization ability of MFRA across multiple instruction formats and grounding requirements, validating the effectiveness of its design under realistic embodied navigation scenarios.

5 Conclusion

In this paper, we propose MFRA, a multi-level fusion and reasoning architecture that enhances the agent's ability to reason over visual observations, language instructions and navigation history for VLN. Our work utilizes a hierarchical fusion mechanism that aggregates multi-level features across multiple modalities, ranging from low-level visual cues to high-level semantic concepts. We further design a reasoning module that leverages fused representations to infer navigation actions through instruction-guided attention and dynamic context integration. We illustrate the good interpretability of MFRA and provide case study in deep insights. Our approach achieves excellent improvement on many VLN tasks, demonstrating that hierarchical fusion and reasoning is a promising direction in improving VLN and Embodied AI. For future work, we will improve our MFRA with larger training data and employ it on VLN in continuous environments.

References

[1] Dong An, Yuankai Qi, Yan Huang, Qi Wu, Liang Wang, and Tieniu Tan. 2021. Neighbor-view Enhanced Model for Vision and Language Navigation. arXiv:2107.07201 [cs.CV] https://arxiv.org/abs/2107.07201

[2] Dong An, Yuankai Qi, Yangguang Li, Yan Huang, Liang Wang, Tieniu Tan, and Jing Shao. 2023. BEVBert: Multimodal Map Pre-training for Language-guided Navigation. arXiv:2212.04385 [cs.CV] https://arxiv.org/abs/2212.04385

[3] Dong An, Hanqing Wang, Wenguan Wang, Zun Wang, Yan Huang, Keji He, and Liang Wang. 2024. ETPNav: Evolving Topological Planning for Vision-Language Navigation in Continuous Environments. arXiv:2304.03047 [cs.CV] https://arxiv.org/abs/2304.03047

[4] Peter Anderson, Qi Wu, Damien Teney, Jake Bruce, Mark Johnson, Niko Sünderhauf, Ian Reid, Stephen Gould, and Anton van den Hengel. 2018. Vision-and-Language Navigation: Interpreting visually-grounded navigation instructions in real environments. arXiv:1711.07280 [cs.CV] https://arxiv.org/abs/1711.07280

[5] Enrico Cancelli et al. 2025. Static and dynamic approaches for Embodied Social Navigation from the perspective of an autonomous agent. (2025).

[6] Devendra Singh Chaplot, Ruslan Salakhutdinov, Abhinav Gupta, and Saurabh Gupta. 2020. Neural Topological SLAM for Visual Navigation. arXiv:2005.12256 [cs.CV] https://arxiv.org/abs/2005.12256

[7] Peihao Chen, Xinyu Sun, Hongyan Zhi, Runhao Zeng, Thomas H. Li, Gaowen Liu, Mingkui Tan, and Chuang Gan. 2023. A^2Nav: Action-Aware Zero-Shot Robot Navigation by Exploiting Vision-and-Language Ability of Foundation Models. arXiv:2308.07997 [cs.CV] https://arxiv.org/abs/2308.07997

[8] Shizhe Chen, Pierre-Louis Guhur, Cordelia Schmid, and Ivan Laptev. 2023. History Aware Multimodal Transformer for Vision-and-Language Navigation. arXiv:2110.13309 [cs.CV] https://arxiv.org/abs/2110.13309

[9] Shizhe Chen, Pierre-Louis Guhur, Makarand Tapaswi, Cordelia Schmid, and Ivan Laptev. 2022. Learning from Unlabeled 3D Environments for Vision-and-Language Navigation. arXiv:2208.11781 [cs.CV] https://arxiv.org/abs/2208.11781

[10] 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. arXiv:2202.11742 [cs.CV] https://arxiv.org/abs/2202.11742

[11] Xinlei Chen and C. Lawrence Zitnick. 2014. Learning a Recurrent Visual Representation for Image Caption Generation. arXiv:1411.5654 [cs.CV] https://arxiv.org/abs/1411.5654

[12] J. C. K. Chow. 2017. DRIFT-FREE INDOOR NAVIGATION USING SIMULTANEOUS LOCALIZATION AND MAPPING OF THE AMBIENT HETEROGENEOUS MAGNETIC FIELD. The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences XLII-2/W7 (Sept. 2017), 339–344. doi:10.5194/isprs-archives-xlii-2-w7-339-2017

[13] J. C. K. Chow, I. Detchev, K. D. Ang, K. Morin, K. Mahadevan, and N. Louie. 2018. ROBOT VISION: CALIBRATION OF WIDE-ANGLE LENS CAMERAS USING COLLINEARITY CONDITION AND K-NEAREST NEIGHBOUR REGRESSION. The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences XLII−1 (Sept. 2018), 93–99. doi:10.5194/isprs-archives-xlii-1-93-2018

[14] Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio. 2014. Empirical Evaluation of Gated Recurrent Neural Networks on Sequence Modeling. arXiv:1412.3555 [cs.NE] https://arxiv.org/abs/1412.3555

[15] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. arXiv:1810.04805 [cs.CL] https://arxiv.org/abs/1810.04805

[16] Leo Feng, Frederick Tung, Hossein Hajimirsadeghi, Mohamed Osama Ahmed, Yoshua Bengio, and Greg Mori. 2024. Attention as an RNN. arXiv preprint arXiv:2405.13956 (2024).

[17] 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. arXiv:1806.02724 [cs.CV] https://arxiv.org/abs/1806.02724

[18] Muraleekrishna Gopinathan. 2025. Toward embodied navigation through vision and language. (2025).

[19] Pierre-Louis Guhur, Makarand Tapaswi, Shizhe Chen, Ivan Laptev, and Cordelia Schmid. 2021. Airbert: In-domain Pretraining for Vision-and-Language Navigation. arXiv:2108.09105 [cs.CV] https://arxiv.org/abs/2108.09105

[20] 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. arXiv:2002.10638 [cs.CV] https://arxiv.org/abs/2002.10638

[21] Yicong Hong, Cristian Rodriguez-Opazo, Yuankai Qi, Qi Wu, and Stephen Gould. 2020. Language and Visual Entity Relationship Graph for Agent Navigation. arXiv:2010.09304 [cs.CV] https://arxiv.org/abs/2010.09304

[22] Yicong Hong, Qi Wu, Yuankai Qi, Cristian Rodriguez-Opazo, and Stephen Gould. 2021. A Recurrent Vision-and-Language BERT for Navigation. arXiv:2011.13922 [cs.CV] https://arxiv.org/abs/2011.13922

[23] Cong Hu, Xiao-Zhong Wei, and Xiao-Jun Wu. 2024. DIRformer: A Novel Image Restoration Approach Based on U-shaped Transformer and Diffusion Models. ACM Trans. Multimedia Comput. Commun. Appl. 21, 2, Article 57 (Dec. 2024), 23 pages. doi:10.1145/3703632

[24] Ayush Jain, Nikolaos Gkanatsios, Ishita Mediratta, and Katerina Fragkiadaki. 2022. Bottom up top down detection transformers for language grounding in images and point clouds. In European Conference on Computer Vision. Springer, 417–433.

[25] Himangshu Kalita, Steven Morad, and Jekan Thangavelautham. 2018. Path Planning and Navigation Inside Off-World Lava Tubes and Caves. arXiv:1803.02818 [cs.RO] https://arxiv.org/abs/1803.02818

[26] 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:2010.07954 [cs.CV] https://arxiv.org/abs/2010.07954

[27] Jiacheng Li, Yujie Lu, Jinpeng Wang, Yuanfang Guo, Weizhi Ma, Zhumin Chen, Jun Ma, and Peng Jiang. 2021. RecBERT: A Pre-trained Language Model for Sequential Recommendation. In Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval. 1885–1889. doi:10.1145/3404835.3462891

[28] Xiangyang Li, Zihan Wang, Jiahao Yang, Yaowei Wang, and Shuqiang Jiang. 2023. KERM: Knowledge Enhanced Reasoning for Vision-and-Language Navigation. arXiv:2303.15796 [cs.CV] https://arxiv.org/abs/2303.15796

[29] Bingqian Lin, Yunshuang Nie, Ziming Wei, Jiaqi Chen, Shikui Ma, Jianhua Han, Hang Xu, Xiaojun Chang, and Xiaodan Liang. 2025. NavCoT: Boosting LLM-Based Vision-and-Language Navigation via Learning Disentangled Reasoning. arXiv:2403.07376 [cs.CV] https://arxiv.org/abs/2403.07376

[30] Hao Liu, Yang Yang, Fumin Shen, Lixin Duan, and Heng Tao Shen. 2016. Recurrent Image Captioner: Describing Images with Spatial-Invariant Transformation and Attention Filtering. arXiv:1612.04949 [cs.CV] https://arxiv.org/abs/1612.04949

[31] Mingyan Liu. 2025. A Unified Virtual Mixture-of-Experts Framework:Enhanced Inference and Hallucination Mitigation in Single-Model System. arXiv:2504.03739 [cs.CL] https://arxiv.org/abs/2504.03739

[32] Yuxing Long, Xiaoqi Li, Wenzhe Cai, and Hao Dong. 2023. Discuss Before Moving: Visual Language Navigation via Multi-expert Discussions. arXiv:2309.11382 [cs.RO] https://arxiv.org/abs/2309.11382

[33] Abhinav Moudgil, Arjun Majumdar, Harsh Agrawal, Stefan Lee, and Dhruv Batra. 2021. SOAT: A Scene- and Object-Aware Transformer for Vision-and-Language Navigation. arXiv:2110.14143 [cs.CV] https://arxiv.org/abs/2110.14143

[34] Yuankai Qi, Zizheng Pan, Shengping Zhang, Anton van den Hengel, and Qi Wu. 2020. Object-and-Action Aware Model for Visual Language Navigation. arXiv:2007.14626 [cs.CL] https://arxiv.org/abs/2007.14626

[35] Yuankai Qi, Qi Wu, Peter Anderson, Xin Wang, William Yang Wang, Chunhua Shen, and Anton van den Hengel. 2020. REVERIE: Remote Embodied Visual Referring Expression in Real Indoor Environments. arXiv:1904.10151 [cs.CV] https://arxiv.org/abs/1904.10151

[36] Jianing Qian, Anastasios Panagopoulos, and Dinesh Jayaraman. 2024. Recasting Generic Pretrained Vision Transformers As Object-Centric Scene Encoders For Manipulation Policies. arXiv:2405.15916 [cs.CV] https://arxiv.org/abs/2405.15916

[37] Yanyuan Qiao, Yuankai Qi, Yicong Hong, Zheng Yu, Peng Wang, and Qi Wu. 2022. HOP: History-and-Order Aware Pre-training for Vision-and-Language Navigation. arXiv:2203.11591 [cs.CV] https://arxiv.org/abs/2203.11591

[38] Yanyuan Qiao, Yuankai Qi, Zheng Yu, Jing Liu, and Qi Wu. 2023. March in Chat: Interactive Prompting for Remote Embodied Referring Expression. arXiv:2308.10141 [cs.CV] https://arxiv.org/abs/2308.10141

[39] Yanyuan Qiao, Zheng Yu, and Qi Wu. 2023. VLN-PETL: Parameter-Efficient Transfer Learning for Vision-and-Language Navigation. arXiv:2308.10172 [cs.CV] https://arxiv.org/abs/2308.10172

[40] 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. PmLR, 8748–8763.

[41] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning Transferable Visual Models From Natural Language Supervision. arXiv:2103.00020 [cs.CV] https://arxiv.org/abs/2103.00020

[42] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. 2016. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. arXiv:1506.01497 [cs.CV] https://arxiv.org/abs/1506.01497

[43] Erick Schmidt, Zachary Ruble, David Akopian, and Daniel J. Pack. 2019. Software-Defined Radio GNSS Instrumentation for Spoofing Mitigation: A Review and a Case Study. IEEE Transactions on Instrumentation and Measurement 68, 8 (Aug. 2019), 2768–2784. doi:10.1109/tim.2018.2869261

[44] Sheng Shen, Liunian Harold Li, Hao Tan, Mohit Bansal, Anna Rohrbach, Kai-Wei Chang, Zhewei Yao, and Kurt Keutzer. 2021. How Much Can CLIP Benefit Vision-and-Language Tasks? arXiv:2107.06383 [cs.CV] https://arxiv.org/abs/2107.06383

[45] Chenglu Sun, Shuo Shen, Wenzhi Tao, Deyi Xue, and Zixia Zhou. 2025. Noise-Resilient Symbolic Regression with Dynamic Gating Reinforcement Learning. arXiv preprint arXiv:2501.01085 (2025).

[46] Chiranjib Sur. 2019. CRUR: Coupled-Recurrent Unit for Unification, Conceptualization and Context Capture for Language Representation – A Generalization of Bi Directional LSTM. arXiv:1911.10132 [cs.CL] https://arxiv.org/abs/1911.10132

[47] Hao Tan and Mohit Bansal. 2019. LXMERT: Learning Cross-Modality Encoder Representations from Transformers. arXiv:1908.07490 [cs.CL] https://arxiv.org/abs/1908.07490

[48] Hao Tan and Mohit Bansal. 2019. Lxmert: Learning cross-modality encoder representations from transformers. *arXiv preprint arXiv:1908.07490* (2019).

[49] Hao Tan, Licheng Yu, and Mohit Bansal. 2019. Learning to Navigate Unseen Environments: Back Translation with Environmental Dropout. arXiv:1904.04195 [cs.CL] https://arxiv.org/abs/1904.04195

[50] Hanqing Wang, Wenguan Wang, Wei Liang, Caiming Xiong, and Jianbing Shen. 2021. Structured Scene Memory for Vision-Language Navigation. arXiv:2103.03454 [cs.CV] https://arxiv.org/abs/2103.03454

[51] 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. arXiv:1811.10092 [cs.CV] https://arxiv.org/abs/1811.10092

[52] Xiaohan Wang, Wenguan Wang, Jiayi Shao, and Yi Yang. 2023. Lana: A Language-Capable Navigator for Instruction Following and Generation. arXiv:2303.08409 [cs.CV] https://arxiv.org/abs/2303.08409

[53] Zun Wang, Jialu Li, Yicong Hong, Yi Wang, Qi Wu, Mohit Bansal, Stephen Gould, Hao Tan, and Yu Qiao. 2023. Scaling Data Generation in Vision-and-Language Navigation. arXiv:2307.15644 [cs.CV] https://arxiv.org/abs/2307.15644

[54] Qi Wu, Chunhua Shen, Anton van den Hengel, Peng Wang, and Anthony Dick. 2016. Image Captioning and Visual Question Answering Based on Attributes and External Knowledge. arXiv:1603.02814 [cs.CV] https://arxiv.org/abs/1603.02814

[55] Zhenda Xie, Zheng Zhang, Yue Cao, Yutong Lin, Jianmin Bao, Zhuliang Yao, Qi Dai, and Han Hu. 2022. SimMIM: A Simple Framework for Masked Image Modeling. arXiv:2111.09886 [cs.CV] https://arxiv.org/abs/2111.09886

[56] Shiyang Yan, Jun Xu, Yuai Liu, and Lin Xu. 2019. HorNet: A Hierarchical Offshoot Recurrent Network for Improving Person Re-ID via Image Captioning. arXiv:1908.04915 [cs.CV] https://arxiv.org/abs/1908.04915

[57] Xinlei Yu, Ahmed Elazab, Ruiquan Ge, Jichao Zhu, Lingyan Zhang, Gangyong Jia, Qing Wu, Xiang Wan, Lihua Li, and Changmiao Wang. 2025. ICH-PRNet: a cross-modal intracerebral haemorrhage prognostic prediction method using joint-attention interaction mechanism. *Neural Networks* 184 (2025), 107096.

[58] Arthur Zhang, Harshit Sikchi, Amy Zhang, and Joydeep Biswas. 2025. CREStE: Scalable Mapless Navigation with Internet Scale Priors and Counterfactual Guidance. arXiv:2503.03921 [cs.RO] https://arxiv.org/abs/2503.03921

[59] Chongyang Zhao, Yuankai Qi, and Qi Wu. 2023. Mind the Gap: Improving Success Rate of Vision-and-Language Navigation by Revisiting Oracle Success Routes. arXiv:2308.03244 [cs.CV] https://arxiv.org/abs/2308.03244

[60] Duo Zheng, Shijia Huang, Lin Zhao, Yiwu Zhong, and Liwei Wang. 2024. Towards Learning a Generalist Model for Embodied Navigation. arXiv:2312.02010 [cs.CV] https://arxiv.org/abs/2312.02010

[61] Gengze Zhou, Yicong Hong, and Qi Wu. 2023. NavGPT: Explicit Reasoning in Vision-and-Language Navigation with Large Language Models. arXiv:2305.16986 [cs.CV] https://arxiv.org/abs/2305.16986

[62] Xinzhe Zhou, Wei Liu, and Yadong Mu. 2021. Rethinking the Spatial Route Prior in Vision-and-Language Navigation. arXiv:2110.05728 [cs.CV] https://arxiv.org/abs/2110.05728

[63] Fengda Zhu, Xiwen Liang, Yi Zhu, Xiaojun Chang, and Xiaodan Liang. 2021. SOON: Scenario Oriented Object Navigation with Graph-based Exploration. arXiv:2103.17138 [cs.CV] https://arxiv.org/abs/2103.17138

[64] Fengda Zhu, Yi Zhu, Xiaojun Chang, and Xiaodan Liang. 2020. Vision-Language Navigation with Self-Supervised Auxiliary Reasoning Tasks. arXiv:1911.07883 [cs.CV] https://arxiv.org/abs/1911.07883

[65] Junyou Zhu, Yanyuan Qiao, Siqi Zhang, Xingjian He, Qi Wu, and Jing Liu. 2024. MiniVLN: Efficient Vision-and-Language Navigation by Progressive Knowledge Distillation. arXiv:2409.18800 [cs.CV] https://arxiv.org/abs/2409.18800

[66] Wanrong Zhu, Yuankai Qi, Pradyumna Narayana, Kazoo Sone, Sugato Basu, Xin Eric Wang, Qi Wu, Miguel Eckstein, and William Yang Wang. 2022. Diagnosing Vision-and-Language Navigation: What Really Matters. arXiv:2103.16561 [cs.CV] https://arxiv.org/abs/2103.16561