Agent Journey Beyond RGB: Hierarchical Semantic-Spatial Representation Enrichment for Vision-and-Language Navigation

Xuesong Zhang[1*], Yunbo Xu[1*], Jia Li[1†], Ruonan Liu[2], Zhenzhen Hu[1]

[1]Hefei University of Technology, Hefei, Chnia
[2]Shanghai Jiao Tong University, Shanghai, China
{xszhang_hfut, xuyunbocn}@mail.hfut.edu.cn, jiali@hfut.edu.cn,
ruonan.liu@sjtu.edu.cn, huzhen.ice@gmail.com

Abstract

Navigating unseen environments based on natural language instructions remains difficult for egocentric agents in Vision-and-Language Navigation (VLN). Intuitively, humans inherently ground concrete semantic knowledge within spatial layouts during indoor navigation. Although previous studies have introduced diverse environmental representations to enhance reasoning, other co-occurrence modalities are often naively concatenated with RGB features, resulting in suboptimal utilization of each modality's distinct contribution. Inspired by this, we propose a hierarchical Semantic Understanding and Spatial Awareness (SUSA) architecture to enable agents to perceive and ground environments at diverse scales. Specifically, the Textual Semantic Understanding (TSU) module supports local action prediction by generating view-level descriptions, thereby capturing fine-grained environmental semantics and narrowing the modality gap between instructions and environments. Complementarily, the Depth-enhanced Spatial Perception (DSP) module incrementally constructs a trajectory-level depth exploration map, providing the agent with a coarse-grained comprehension of the global spatial layout. Extensive experiments demonstrate that SUSA's hierarchical representation enrichment not only boosts the navigation performance of the baseline on discrete VLN benchmarks (REVERIE, R2R, and SOON), but also exhibits superior generalization to the continuous R2R-CE.

Code — https://github.com/HCI-LMC/VLN-SUSA

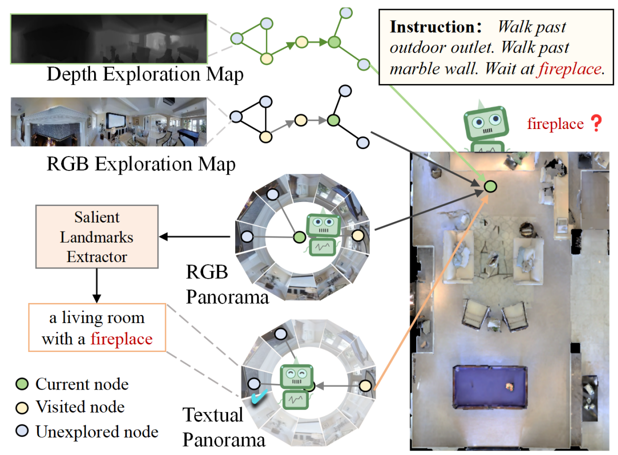

Figure 1: An overview of the proposed SUSA. Beyond RGB inputs, we introduce view-level *textual panoramas* and trajectory-level *depth exploration maps*, supporting the agent's explicit understanding of the environment.

1 Introduction

With the advancement of multimodal technology (Radford et al. 2021; Li et al. 2023a; Zhang et al. 2024) in the artificial intelligence community, embodied VLN tasks (Krantz et al. 2020; Qi et al. 2020; Das et al. 2018) have garnered significant attention due to their promising applications in robotics and intelligent assistance. Conventional VLN tasks (Anderson et al. 2018; Krantz et al. 2020) focused on step-by-step navigation in unseen environments based on detailed instructions, while goal-oriented VLN tasks (Zhu et al. 2021; Qi et al. 2020) demand agents to identify predefined objects based on high-level instructions. Consequently, VLN agents are required to profoundly cognize the environmental spatial information and ground them with textual instructions to brilliantly accomplish navigation tasks.

Recently, increasing the diversity and quantity of augmented environments (Wang et al. 2023b,c; Liu, Wang, and Yang 2024) has emerged as a popular strategy to facilitate the generalization ability of VLN agents in unseen scenarios. This trend is largely motivated by the limited navigational environments in simulators, as well as the powerful fitting capacity of Transformer-based VLN models (Chen et al. 2021; Pashevich, Schmid, and Sun 2021). Nonetheless, VLN agents still struggle to accurately ground the landmarks mentioned in language instructions with the RGB-based visual scene, due to the intrinsic modality heterogeneity. To address this, recent studies introduce textual semantics as perceptual representations to guide navigation (Pan et al. 2024). These approaches leverage candidate object concepts (Wang et al. 2023a; Lin et al. 2023), large language models (LLMs) (Lin et al. 2024), or external knowledge (Li et al. 2023b) to explicitly reason about textual descriptions of landmarks and enhance environmental grounding. However, these methods tend to combine textual and visual rep-

[*]These authors contributed equally.
[†]Corresponding author.

resentations in a straightforward manner, which limits the effective utilization of each modality and constrains the interpretability of their individual contributions. Similarly, co-occurring environmental cues from other modalities—such as frequency (He et al. 2024), spectrum (Hwang et al. 2023), and depth (Tan et al. 2022)—have also been preliminarily explored, though remain underutilized.

In this work, we propose the hierarchical Semantic Understanding and Spatial Awareness (SUSA) architecture, which augments the agent's environmental comprehension by effectively perceiving and exploiting fine-grained textual semantics and coarse-grained depth-aware spatial information. To enable fine-grained semantic grounding, as shown in Fig. 1, we construct a view-level textual panorama by extracting salient landmarks as textual semantics, thereby bridging the gap between visual and linguistic modalities. The most instruction-relevant view is then selected for local navigation prediction. For global action prediction, we progressively construct a trajectory-level depth exploration map to strengthen the agent's coarse-grained spatial perception. Finally, SUSA hierarchically aggregates these hybrid semantic-spatial representations to perceive complementary environmental information and employs contrastive learning to align them with natural language instructions. Extensive experiments on four VLN benchmarks (R2R (Anderson et al. 2018), REVERIE (Qi et al. 2020), SOON (Zhu et al. 2021) and R2R-CE (Krantz et al. 2020)) demonstrated the superiority of hybrid semantic-spatial representations enrichment. The contributions of our SUSA are as follows:

- SUSA devises a textual-aware semantic understanding (TSU) module to perceive fine-grained textual semantics of environments, and explicitly address modality heterogeneity between instructions and environments.
- SUSA further constructs a depth exploration map within the depth-enhanced spatial perception (DSP) module to complementarily perceive global environmental layouts and reinforce holistic spatial awareness beyond RGB.
- Experiments demonstrate that SUSA substantially improves upon the baseline, achieving state-of-the-art performance on three discrete VLN tasks, and exhibits promising generalization in continuous environments.

2 Related Work

Vision-and-Language Navigation (VLN). Vision-and-Language Navigation (VLN) tasks (Anderson et al. 2018; Qi et al. 2020; Zhu et al. 2021) require an agent to navigate in embodied environments sequentially towards a designated location or object based on natural language instructions. Given that the VLN task can be formulated as a partially observable Markov decision process, early methods (Wang et al. 2019; Fried et al. 2018) primarily relied on reinforcement learning and imitation learning paradigms to improve navigation performance, incorporating various strategies and auxiliary loss (Wang et al. 2019) designs. More recently, to better model the relationship between the environment and instructions, many researchers (Chen et al. 2021; Pashevich, Schmid, and Sun 2021) have focused on transformer-based architectures, which effectively model interactions among instructions, environmental features, and historical navigation trajectories.

Environment Representations in VLN. Limited training sources and environmental representations in VLN tasks hinder the agent to generalize the unknown layouts. To alleviate this limitation, recent endeavors (Wang et al. 2023b; Chen et al. 2022a) augment the quantity of environmental scenes, while others diversify environmental representations by constructing graphs (Georgakis et al. 2022; Huang et al. 2023), grids (Wang et al. 2023c; An et al. 2023), or volumetric environments (Liu, Wang, and Yang 2024). Beyond RGB inputs, recent efforts have explored other co-occurrence multimodal(Li et al. 2023b; He et al. 2024) cues to enrich environmental representations. For instance, SEAT (Wang et al. 2024b) enhances cross-domain environmental perception by querying the Matterport3D simulator (Chang et al. 2017) for depth and semantics. However, SEAT fuses these multi-type representations into a unified space without explicitly grounding each modality. Moreover, these semantic representations (Wang et al. 2024b; Hong et al. 2023) are not inherently textual, limiting their ability to bridge the modality gap. Other approaches incorporate depth information (An et al. 2023; Wang et al. 2023c); however, they typically project depth values onto the spatial coordinates of RGB images without independently grounding the depth modality. In contrast, our method constructs the view-level textual panoramas and trajectory-level depth exploration maps, individually grounding them with instructions to perceive rich environmental representations.

3 Methodology

3.1 Preliminary

Task Formulation. We focus on discrete VLN tasks (Anderson et al. 2018; Qi et al. 2020; Zhu et al. 2021) within a Matterport3D simulator (Chang et al. 2017), modeled as an undirected graph $\mathcal{G} = \{\mathcal{N}, \mathcal{E}\}$, where $\mathcal{N}$ denotes nodes and $\mathcal{E}$ signifies the connectivity paths. At the outset of each episode, the agent equipped with a GPS sensor, RGB and depth cameras, is situated at a randomly navigable node and receives a language instruction $I = \{\mathrm{w}_i\}_{i=1}^{l}$ including l words. We use BERT (Devlin et al. 2019) to extract instruction features and then embed them as $\mathcal{I} = \{w_1, \dots, w_l\}$. At each time step t, the agent can perceive the surrounding RGB panorama $R_t = \{\mathrm{r}_{t,i}\}_{i=1}^{n=36}$ and depth panorama $D_t = \{\mathrm{d}_{t,i}\}_{i=1}^{n=36}$ along with n corresponding views. During navigation, the agent makes the next action a_t by selecting one navigable node from the candidate actions $\mathcal{A}_t = \{a_{t,i}\}_{i=0}^{k}$.

Overview of the Proposed Method. Our objective is to enrich complementary environmental content by equipping the agent with a textual semantic panorama and a depth exploration map. Following prevailing approaches (He et al. 2024), we adopt DUET (Chen et al. 2022b) as our baseline, which merely utilizes RGB panoramas and topological maps as environmental representations for both local and global action predictions. The SUSA framework consists of three main modules: (a) *TSU*: textual-aware semantic selection is detailed in Section 3.2, (b) *DSP*: depth-enhanced visual en-

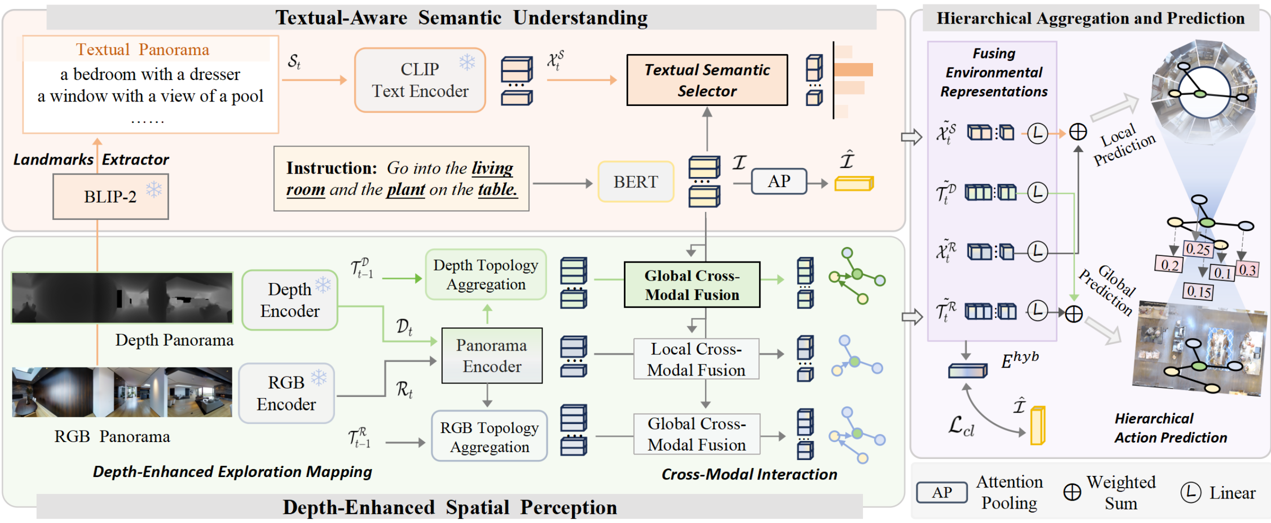

Figure 2: The detailed architecture of our proposed SUSA model. The orange and green arrows highlight our proposed Textual-Aware Semantic Understanding (*TSU*) and Depth-Aware Spatial Perception (*DSP*) modules, respectively. The Hierarchical Aggregation and Prediction (*HAP*) module is designed for hierarchical aligning environmental representations with instructions.

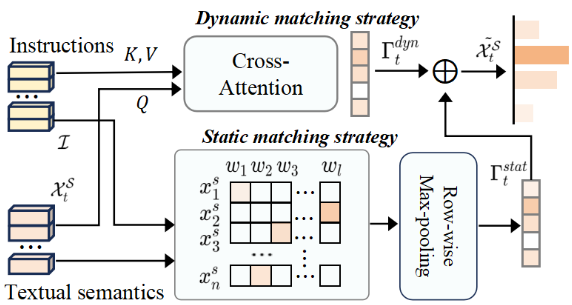

Figure 3: Illustration of the *TSU* module in Fig.2, which matches instructions and semantic features by static and dynamic matching strategies.

vironment exploration is presented in Section 3.3, and (c) *HAP*: hybrid representation aggregation and hierarchical action prediction is subsequently introduced in Section 3.4, as illustrated in Fig. 2.

3.2 Textual-Aware Semantic Understanding

In this module, we build a textual semantic panorama that enables the agent to understand the environment from a textual perspective, allowing it to select the most relevant navigable view as the optimal local navigation action.

Salient Landmarks Extractor. The pretrained visual-language model readily provides the capability for agents to perceive textual semantics by generating detailed captions from RGB images. As illustrated in Fig. 2, we utilize off-the-shelf BLIP-2-FlanT5-xxL (Li et al. 2023a), a powerful visual-language model, to discretize R_t and generate corresponding textual descriptions for each view using the simple prompt “a photo of”. Following (Gao et al. 2024; Li and Bansal 2023), this process yields a textual panorama $\mathcal{S}_t = \{s_{t,i}\}_{i=1}^{n=36}$, where each view $s_{t,i}$ contains informed descriptions of salient landmarks. These descriptions typically involve room types and objects (e.g.,“a living room with two chairs”), which assist in the object recognition. We further discuss the quality of the generated descriptions in supplementary materials. Additionally, we employ the CLIP (ViT-L-14-336px) text encoder (Radford et al. 2021) to extract textual semantic panorama features, embedding them as $\mathcal{X}_t^{\mathcal{S}} = \{x_{t,1}^s, \dots, x_{t,n}^s\}$. Thanks to the powerful yet frozen BLIP-2 and CLIP models, we empower the agent to obtain textual semantic features from the navigable nodes in its vicinity during navigation.

Textual Semantic Selector. Similar to (Pan et al. 2024), we treat the textual semantic as a perceptual representation for navigation in this module, directly matching salient landmarks with the instructions. As shown in Fig. 2 and Fig. 3, our textual semantic selector combines both *static* and *dynamic* matching strategies to select the most instruction-relevant view in the textual panorama $\mathcal{X}_t^{\mathcal{S}}$.

Static Matching: The most straightforward approach to semantic alignment is by calculating similarity, where higher similarity indicates stronger relevance between the textual semantics of each view and the instruction. Consequently, we calculate the cosine similarity matrix $\mathbf{M} \in \mathbb{R}^{n\times l}$, which represents the correlation between each word in the instruction $\mathcal{I}$ and each view in the textual panorama $\mathcal{X}_t^{\mathcal{S}}$. Nevertheless, the textual semantics generated only correspond to salient landmarks (e.g.,“plant,” or “table”) rather than actions (e.g., “go to” or “ turn left”) in the instructions. Therefore, we then apply row-wise max-pooling to purify the static instruction relevance $\Gamma_t^{stat} \in \mathbb{R}^n$ to match the most relevant landmarks between textual semantics and instructions at step t :

$$\mathbf{M}_{i,j} = \frac{\mathcal{X}_{t,i}^{\mathcal{S}} \cdot \mathcal{I}_j}{\|\mathcal{X}_{t,i}^{\mathcal{S}}\|\|\mathcal{I}_j\|}, \quad \Gamma_t^{stat} = \max_j \mathbf{M}_{i,j} \tag{1}$$

where $i = 1, \ldots, n$ and $j = 1, \ldots, l$. Although straightforward, the static strategy relies on the monotonic similarity matrix to match textual semantics $\mathcal{X}_t^{\mathcal{S}}$ and instructions $\mathcal{I}$, struggling with long-term navigation reasoning.

Dynamic Matching: Furthermore, a dynamic matching strategy is introduced to better capture the intricate relationships between the environment and instruction. Specifically, we utilize a standard multi-layer transformer decoder to dynamically match the relevance Γ_t^{dyn} between textual semantics and instructions, as formalized by:

$$\Gamma_t^{dyn} = \text{TCA}(\mathcal{X}_t^{\mathcal{S}}, \mathcal{I}, \mathcal{I}), \tag{2}$$

where TCA denotes a **t**extual modality **c**ross-**a**ttention mechanism that models textual semantic-instruction relationships. TCA comprises a cross-attention layer, a feed-forward network (FFN) layer, and LayerNorm, enabling the agent to adapt to long-term navigation sequences.

To trade off computational efficiency and navigation accuracy, we flexibly combine the relevance of static Γ_t^{sim} and dynamic Γ_t^{att}. The latent semantic panoramic feature $\tilde{\mathcal{X}_t^{S}}$ is expressed as follows:

$$\tilde{\mathcal{X}_t^{S}} = \delta * \Gamma_t^{stat} + (1 - \delta) * \Gamma_t^{dyn}, \tag{3}$$

where the balance factor δ regulates the relative contributions of each matching strategy. By mapping the environment to the same textual modality, the textual semantic selector enables the agent to explicitly understand the environment context from a textual semantic perspective, thus facilitating grounded language instructions.

3.3 Depth-Enhanced Spatial Perception

Depth images, which also contain more intuitive and easily distinguishable structural information, were not adequately exploited in earlier works (Tan et al. 2022). To enhance the spatial perception, we concurrently construct a depth-based exploration map to memorize depth-based navigation trajectory alongside the RGB-based one (Chen et al. 2022b).

Depth-Enhanced Exploration Mapping. Concretely, we obtain ground truth depth images for each view from the Matterport3D simulator. As illustrated in Fig. 2, we then employ ResNet-50 (He et al. 2016) (pretrained on the Gibson (Xia et al. 2018)) and CLIP (ViT-L/14@336px) as the depth and RGB encoders, respectively, to extract panoramic depth features $\mathcal{D}_t \in \mathbb{R}^{n \times d_v}$ and panoramic RGB features $\mathcal{R}_t \in \mathbb{R}^{n \times d_v}$, where d_v represents the dimension of each view. Moreover, we employ a two-layer transformer-based panorama encoding module to perform self-attention and sequentially encode the depth and RGB panoramic images. The panorama encoding modules for both modalities share weights, since each view in the depth and RGB panoramas is accordant. Subsequently, we aggregate each visited node and its adjacent views by average pooling to construct depth and RGB exploration maps in parallel. The explored depth and RGB trajectory maps are formalized as $\mathcal{T}_t^{\mathcal{D}} \in \mathbb{R}^{k \times d_n}$ and $\mathcal{T}_t^{\mathcal{R}} \in \mathbb{R}_t^{k \times d_n}$. Here, k denotes the number of navigable nodes at step t, while d_n represents the dimension of each navigable node. These exploration maps memorize the agent navigation trajectory information for global action predictions, which serve to guide the agent traversal across various navigable nodes when considering backtracking.

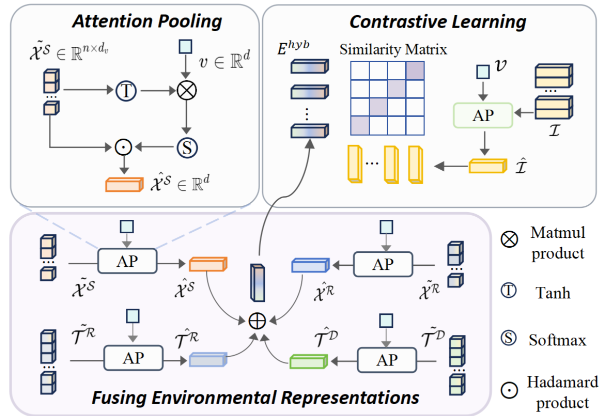

Figure 4: The pipeline for fusing hybrid environmental representations in the proposed *HAP* module in Fig. 2.

Cross-Modal Interaction and Reasoning. To defer to instructions, we utilize a multi-layer cross-encoder that interacts with both the language instructions and the depth exploration map features to generate the refined depth map $\tilde{\mathcal{T}^{\mathcal{D}}}$. This encoder comprises four residual connection layers of a cross-modal transformer, each incorporating graph-aware self-attention (GASA) (Chen et al. 2022b), cross-modal attention, and a feed-forward network. The GASA operation is formally expressed as $\text{GASA}(X) = \eth(X\Theta_q(X\Theta_k)^T + \mathcal{M})X\Theta_v$ where $\mathcal{M}$ represents a learnable distance matrix encoding the relative distances between navigation nodes, $\eth$ denotes the Softmax activation function, and $\Theta_q, \Theta_k, \Theta_v$ are learnable weight matrices. In parallel, we facilitate similar cross-modal interactions between the language instructions and the RGB exploration map $\mathcal{T}^{\mathcal{R}}$, as well as the RGB panoramas $\mathcal{X}^{\mathcal{R}}$. These are encoded as $\tilde{\mathcal{T}^{\mathcal{R}}}$ and $\tilde{\mathcal{X}^{\mathcal{R}}}$, thereby enabling the model to perform both global and local predictions based on the given instructions $\mathcal{I}$. Compared to the standalone $\mathcal{T}^{\mathcal{R}}$, the incorporation of a depth exploration map $\mathcal{T}^{\mathcal{D}}$ enhances the perceptual clarity of environmental structures, mitigating the risk of overfitting to RGB-based visual noises (Ilinykh, Emampoor, and Dobnik 2022) or biases (Hu et al. 2019; Wang et al. 2024a).

3.4 Hierarchical Aggregation and Prediction

Building on aforementioned methods, at the current state t, the agent synchronously perceives the environment through multiple modalities—textual semantic, depth, and RGB —at various scales (local and global). To encourage the agent to aggregate complementary context from these hybrid representations, we encode these features into a low-dimensional latent space via attention pooling and apply contrastive learning to align them with the instructions.

Hybrid Representations Aggregation. As depicted in Fig. 4, we employ attention pooling to project all environmental representations at step t, along with the corresponding instruction, into an equivalent dimension d. Specifically, the attention pooling AP operation is defined as follows:

$$\hat{\mathcal{H}} = \mathrm{AP}(\mathcal{H}, v), \quad \mathcal{H} \in \{\tilde{\mathcal{X}^{\mathcal{S}}}, \tilde{\mathcal{T}^{\mathcal{D}}}, \tilde{\mathcal{X}^{R}}, \tilde{\mathcal{T}^{R}}, \mathcal{I}\}, \tag{4}$$

where $v \in \mathbb{R}^d$ denoting the learnable query vector and $\hat{\mathcal{H}} \in \mathbb{R}^d$. The AP operator is given by:

$$\eta = \eth\left(\tau(\mathcal{H}) \otimes v^\top\right), \quad \hat{\mathcal{H}} = \tau\left(\eta \odot \mathcal{H}\right), \tag{5}$$

with τ representing the tanh activation function. Subsequently, we perform weighted aggregation of the hybrid environmental representations E^{hyb}, as follows:

$$\mathcal{B} = \begin{bmatrix}\beta_1 & \beta_2 & \beta_3 & \beta_4\end{bmatrix}^\top = \sigma(\mathrm{FFN}[\tilde{\mathcal{X}_0^{\mathcal{S}}}; \tilde{\mathcal{X}_0^{\mathcal{R}}}; \tilde{\mathcal{T}_0^{\mathcal{D}}}; \tilde{\mathcal{T}_0^{\mathcal{R}}}]) \tag{6}$$

$$E^{\mathrm{hyb}} = \begin{bmatrix}\hat{\mathcal{X}^{\mathcal{S}}} & \hat{\mathcal{X}^{\mathcal{R}}} & \hat{\mathcal{T}^{\mathcal{D}}} & \hat{\mathcal{T}^{\mathcal{R}}}\end{bmatrix} \cdot \mathcal{B}, \tag{7}$$

where σ stands for the sigmoid activation function, and $[;]$ denotes concatenation.

Hierarchical Action Prediction. We perform hierarchical action prediction by synergistically fusing the prediction scores from multiple branches to capture complementary information. Specifically, the scores for each branch are computed via separate FFN: $P_{\bar{\mathcal{H}}} = \mathrm{FFN}_{\bar{\mathcal{H}}}(\bar{\mathcal{H}})$, where $\bar{\mathcal{H}} = \{x \in \mathcal{H} \mid x \neq \mathcal{I}\}$. Then, we apply weighted fusion to the textual semantic panorama $\mathcal{X}^{\mathcal{S}}$ and the RGB panorama $\mathcal{X}^{\mathcal{R}}$ to make local predictions, along with global predictions derived from exploration maps based on depth $\mathcal{T}^{\mathcal{D}}$ and RGB $\mathcal{T}^{\mathcal{R}}$, which is formulated as:

$$P_{\mathcal{X}} = \beta_1 P_{\mathcal{X}^{\mathcal{S}}} + \beta_2 P_{\mathcal{X}^{\mathcal{R}}}, \tag{8}$$

$$P_{\mathcal{T}} = \beta_3 P_{\mathcal{T}^{\mathcal{D}}} + \beta_4 P_{\mathcal{T}^{\mathcal{R}}}. \tag{9}$$

We convert local predictions into global actions and predict a_t via dynamic fusing (Chen et al. 2022b) $P_{\mathcal{T}}$ and $P_{\mathcal{X}}$.

3.5 Training

Partially Pretraining. Our pre-training phase followed prior work (Chen et al. 2022b), utilizing auxiliary tasks such as Masked Language Modeling (MLM), Masked Region Classification (MRC), and Single-step Action Prediction (SAP). Object Grounding (OG) task to localize target objects in REVERIE and SOON. Given the sparsity of depth images and textual semantics compared to RGB images, completely pre-training all parameter of SUSA may lead to overfitting and increased training costs. As shown in Fig. 2, only the computation flows indicated by black arrows were pre-trained. The effectiveness of this partial pre-trained strategy is demonstrated in Section 4.3.

Contrastive and Imitation Learning. During fine-tuning, we employ a contrastive learning loss $\mathcal{L}_{cl}$ to align the hybrid environmental representations E^{hyb} with the corresponding instructions $\mathcal{I}$:

$$\begin{aligned}\mathcal{L}_{cl} = &-\tfrac{1}{2B}\textstyle\sum_{i=1}^{B} \log \frac{\exp((E_i^{\mathrm{hyb}}, \mathcal{I}_i)/t)}{\sum_{j=1}^{B}\exp((E_i^{\mathrm{hyb}}, \mathcal{I}_j)/t)} \\ &-\tfrac{1}{2B}\textstyle\sum_{i=1}^{B} \log \frac{\exp((\mathcal{I}_i, E_i^{\mathrm{hyb}})/t)}{\sum_{j=1}^{B}\exp((\mathcal{I}_i, E_j^{\mathrm{hyb}})/t)},\end{aligned} \tag{10}$$

where B and t represent the batch size and the temperature. Then, we train the model based on hybrid environment representations by performing single-step action prediction (Chen et al. 2021, 2022b) with a contrastive learning loss $\mathcal{L}_{cl}$. The hybrid single action prediction loss in imitation learning is computed as $\mathcal{L}_{hsap}^{gt} = \sum_{t=1}^{T} -\log p(a_t^{gt}|\mathcal{I}, \mathcal{S}_t, \mathcal{D}_{<t}, \mathcal{R}_{<t})$ and $\mathcal{L}_{hsap}^{\star} = \sum_{t=1}^{T} -\log p(a_t^{\star}|\mathcal{I}, \mathcal{S}_t, \mathcal{D}_{<t}, \mathcal{R}_{<t})$, where a_t^{gt} represents the ground truth action, and $a_t^{\star}$ denotes the pseudo-target action sampled from current exploration map. The final training loss is then expressed as follows:

$$\mathcal{L}_{SUSA} = \lambda_1 \mathcal{L}_{hsap}^{gt} + \mathcal{L}_{hsap}^{\star} + \lambda_2 \mathcal{L}_{cl} \tag{11}$$

where λ_1 and λ_2 denote loss weights.

4 Experiments

4.1 Task Setup and Implementation

Datasets. We mainly evaluate our model on three diverse VLN benchmarks. **R2R** (Anderson et al. 2018) focuses solely on navigation following detailed instructions. **REVERIE** (Qi et al. 2020) requires the agent to recognize the correct object from candidate bounding boxes upon reaching the navigation goal. **SOON** (Zhu et al. 2021) challenges agents to generate candidate bounding boxes using an object detector. Given that discrete agents can serve as planners of ETPNav (An et al. 2024), we also evaluate SUSA on the continuous R2R-CE (Krantz et al. 2020). **R2R-CE** is constructed from the discrete Matterport3D and executed in the continuous Habitat simulator (Savva et al. 2019).

Metrics. We evaluate the navigation performance using standard metrics: Success Rate (SR), Oracle Success Rate (OSR), Navigation Error (NE), and Success Rate weighted by Path Length (SPL). For the REVERIE and SOON tasks, which involve object identification, we also evaluate object grounding metrics: Remote Grounding Success (RGS) and RGS weighted by Path Length (RGSPL). Given that **SPL** optimally balances navigation success rate and trajectory length, we employ it as our primary metric.

Implementation Details. Pre-training was performed for 100k iterations with a batch size of 32. Fine-tuning was performed for 25k iterations for VLN tasks. For fine-tuning, we used batch sizes of 4, 8, 2 and 16 for R2R, REVERIE, SOON and R2R-CE, respectively. The default hyperparameters are $\lambda_1 = 0.2$ and $\lambda_2 = 0.8$. The node and view dimensions are identical: $d_v = d_n = d = 768$. We minimized changes to the baseline (DUET) (Chen et al. 2022b) settings and used no additional annotations. All experiments were conducted on a **single** NVIDIA RTX 4090 GPU.

4.2 Comparisons with State of the Art

We primarily conduct a comprehensive comparison on three discrete VLN tasks—REVERIE (Qi et al. 2020), R2R (Anderson et al. 2018), and SOON (Zhu et al. 2021). For fairness, we exclude methods that involve pre-exploration (Sigurdsson et al. 2023) or utilize large-scale data augmentation (Wang et al. 2023b). Tables 1 and 2 present a comparative analysis on REVERIE, R2R, and SOON. Our method achieves, and in some cases surpasses, the performance of previous state-of-the-art approaches on various metrics. For instance, SUSA achieves an SPL/RGSPL of 41.5%/27.3% on the REVERIE test unseen split, significantly outperforming the baseline (DUET) by a large margin of 5.5%/5.3%.

Methods	REVERIE						R2R					
	Val Unseen			Test Unseen			Val Unseen			Test Unseen		
	SR	**SPL**	RGSPL	SR	**SPL**	RGSPL	SR	**SPL**	NE↓	SR	**SPL**	NE↓
RCM (Wang et al. 2019)	9.2	6.9	3.8	7.8	6.6	3.1	43	-	6.0	43	38	6.1
GridMM (Wang et al. 2023c)	51.3	36.4	24.5	53.1	36.6	23.4	**75**	64	-	**73**	62	-
KERM (Li et al. 2023b)	49.0	34.8	24.1	52.2	37.4	23.1	71.9	60.9	3.2	69.7	59.2	3.6
AZHP (Zhan et al. 2024)	49.0	36.2	24.1	52.5	36.1	22.5	71	60	3.2	69	59	3.4
FDA (He et al. 2024)	47.5	35.9	24.3	49.6	36.4	22.0	72	64	3.4	69	62	3.41
CONSOLE (Lin et al. 2024)	50.0	34.4	23.3	**55.1**	37.1	22.2	73	63	3.0	72	61	3.3
ACME (Wu et al. 2025)	49.5	32.4	24.0	51.9	34.7	23.6	72.8	62.3	3.12	70.4	61.2	3.68
DUET (Chen et al. 2022b)	46.9	33.7	23.0	52.5	36.0	22.0	72	60	3.3	69	59	3.65
SUSA (Ours)	**51.7**	**38.8**	**26.5**	54.3	**41.5**	**27.3**	73.0	**64.8**	**3.0**	72.5	**63.8**	**3.20**

Table 1: Comparison with the state of the art on REVERIE and R2R. **Bold** highlights the best performance in each column. ↓ indicates better performance with lower values. "–": unavailable statistics. % is omitted for brevity.

Method	Val Unseen		Test Unseen	
	SPL	RGSPL	**SPL**	RGSPL
GridMM (Wang et al. 2023c)	24.8	3.9	21.2	4.1
KERM (Li et al. 2023b)	23.1	4.4	-	-
SEAT (Wang et al. 2024b)	24.8	3.9	22.5	4.4
ACME (Wu et al. 2025)	27.8	4.3	22.3	5.5
DUET (Chen et al. 2022b)	22.5	3.7	21.4	4.1
SUSA(Ours)	**30.8**	**6.4**	**25.4**	**5.9**

Table 2: Comparison with the state of the art on the SOON.

Methods	Val Unseen		Test Unseen	
	SR	**SPL**	SR	**SPL**
GridMM (Wang et al. 2023c)	49	41	46	39
AO-Planner (Chen et al. 2025)	47	33	-	-
ETPNav (An et al. 2024)	57	49	55	48
DUET (Chen et al. 2022b)	51.9	41.2	47.1	40.5
SUSA (Ours)	<u>52.7</u>	<u>43.6</u>	<u>50.9</u>	<u>43.9</u>

Table 3: Comparison with the state of the art on R2R-CE. We directly transfer the DUET and SUSA as planners of ETPNav into the continuous environment without pretraining. <u>Underline</u> indicates improvements over the baseline.

Furthermore, we briefly transfer the agent into continuous environments to coarsely assess its adaptability and generalization. Results in Table 3 indicate that SUSA achieves superior performance over the baseline in the continuous R2R-CE task, highlighting the potential of hierarchical representation enrichment in continuous navigation scenarios.

4.3 Diagnostic Experiment

1) Image Features vs. Structural Features. We analyze RGB and depth features contributions within the DSP module (Table 4). For RGB images, substituting ViT (Dosovitskiy et al. 2020) features with those from CLIP (Radford et al. 2021) yields marked improvements across all metrics. As for depth features, we leveraged depth features extracted by ResNet-50 (He et al. 2016) pre-trained on the ImageNet (Deng et al. 2009) or Gibson (Xia et al. 2018) datasets to investigated the effect of depth exploration maps. Compared to ImageNet, the Gibson simulator provides spatial structural information that serves as valuable navigation priors. Therefore, the ResNet-50 pre-trained on the Gibson yields notable improvements in the SPL metric without compromising SR or RGS. These improvements indicate that better spatial perception enhances navigation efficiency.

RGB	Depth	SR	**SPL**	RGS	RGSPL
ViT	w/o	46.9	33.7	32.1	23.0
CLIP	w/o	51.5	35.7	**35.1**	24.4
CLIP	ImageNet	50.1	34.6	34.3	23.6
CLIP	Gibson	**52.0**	**39.2**	35.0	**26.5**

Table 4: Ablation of different visual representations on the REVERIE validation unseen split. Gibson and ImageNet serve as pre-training sources for the depth encoder (ResNet-50) in the DSP module.

δ	REVERIE			R2R		
	SR	**SPL**	RGSPL	SR	**SPL**	nDTW
w/o	52.0	39.2	26.5	72.8	62.7	69.3
0	53.9	38.8	**26.8**	72.2	63.7	70.7
0.5	**55.0**	**39.5**	26.7	**73.0**	**64.8**	**71.0**
1.0	53.5	37.8	25.9	72.3	62.9	69.1
adaptive	52.4	38.2	26.2	71.9	63.1	69.9

Table 5: Ablation of static and dynamic matching strategies in the TSU module, balanced by the balance factor δ.

2) Static Matching vs. Dynamic Matching. To assess the impact of static matching ($\delta = 0$) and dynamic matching ($\delta = 1$) strategies in the TSU module, we perform ablation studies based on balance factor δ on the R2R and REVERIE. Table 5 summarizes these results, where *w/o* indicates the exclusion of the TSU module. Furthermore, *adaptive* denotes that δ is a learnable parameter which adaptively adjusts

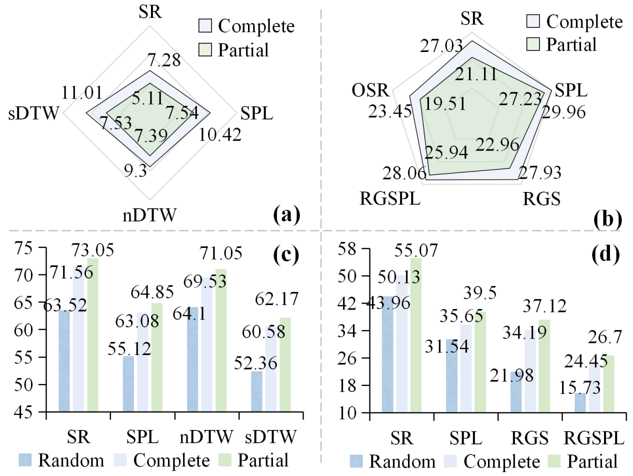

Figure 5: (a) and (b) show the performance gap (↓, lower values indicate better performance.) between seen and unseen environments, while (c) and (d) present key metrics (↑) for different pretraining strategies on the R2R and REVERIE.

ID	**DSP**	**TSU**	**HAP**	SR	SPL	RGSPL
1	✗	✗	✗	46.9|52.5	33.7|36.0	23.0|22.0
2	✓	✗	✗	52.0|52.5	39.2|39.5	26.5|24.9
3	✗	✓	✗	53.1|53.8	38.3|40.8	26.7|26.4
4	✓	✓	✗	55.0|51.9	39.5|37.8	26.7|25.2
5	✓	✓	✓	51.7|54.3	38.8|41.5	26.5|27.3

Table 6: Ablation study on different components of SUSA on the REVERIE (validation|test) unseen split.

the contribution of each matching strategy. While the TSU module can achieve superior performance under the *adaptive* setting, it underperforms when $\delta = 0.5$. Specifically, at $\delta = 0.5$, the SPL reaches 39.50% and 64.85% on the two datasets. Excellent results demonstrate dual merits of TSU module, which boosts object identification accuracy on the REVERIE while simultaneously improving the fidelity between navigation trajectories and instructions on the R2R.

3) Improved Generalization via Pretraining. As aforementioned in Section 3.5, pretraining the entire SUSA framework on multiple auxiliary tasks may lead to overfitting, which negatively impact the generalization in unseen environments. As shown in Fig. 5, the partial pretraining strategy not only consistently narrows the performance gap between seen and unseen environments but also outperforms complete/random pretraining strategies in navigation performance. In Fig. 5 (a), the partial pretraining model achieves an SR gap of only 21.11% between the seen and unseen splits on REVERIE, reducing the gap by 5.92% compared to the complete pretraining strategy. Additionally, as depicted in Fig. 5 (c), multiple metrics indicate that partial pretraining results in superior performance. Fig. 5 (b) and (d) illustrate similar trends are observed on R2R, further emphasizing the generalization of the partial pretraining strategy.

4) Overall Design. To thoroughly evaluate the efficacy of key components, we conduct ablation experiments on

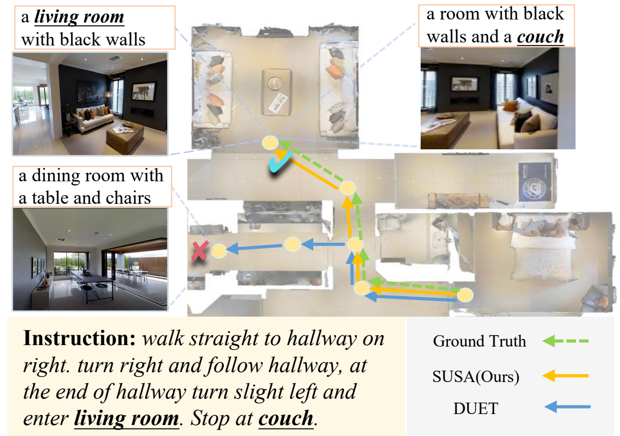

Figure 6: Predicted trajectories of SUSA and DUET on the R2R validation unseen split.

the REVERIE validation and test unseen splits in Table 6. An intriguing phenomenon is observed in Table 6: directly incorporating the DSP and TSU modules (#4) results in an improvement on validation unseen (e.g., SR=55.07, SPL=39.50), while performance on test unseen deteriorates. Conversely, contrastive learning within the HAP module (#5) produces opposite outcomes, particularly with respect to the SR metric. This may be attributed to the learnable token v, which, while pooling hybrid environmental representations, also introduces slight noise that causes fluctuations of navigation performance. Nevertheless, the overall performance is progressively improved with the inclusion of DSP, TSU and HAP modules.

4.4 Qualitative Analysis

As shown in Fig. 6, compared to DUET, our SUSA agent is able to accurately stop near the "living room with a couch." In contrast, the DUET agent, which relies exclusively on RGB environment representations, is prone to making elusive actions, hindering its ability to accurately follow instructions and reach the target location. This manifests that, thanks to the salient landmarks provided by the SUSA architecture, the agent is better able to navigate to the target location by grounding textual instructions.

5 Conclusion

We present the SUSA architecture, which hierarchical exploits semantic-spatial representations beyond RGB to facilitate environmental understanding and instruction grounding. Specifically, we propose the TSU module, enabling the agent to identify the most instruction-relevant textual semantic. The DSP module complementarily enhances spatial awareness by incrementally constructing and grounding depth exploration maps. Experimental results on the R2R, REVERIE, SOON and R2R-CE benchmarks demonstrate the navigational superiority of our approach. Benefiting from hierarchical representation enrichment and alignment, SUSA exhibits excellent generalization performance on unseen environments and modality interpretability.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (NSFC) under Grants Nos. 62202139, 62573292 and 62172138, and also supported by the Fundamental Research Funds for the Central Universities under Grants Nos.JZ2025HGTB0226, JZ2024HGTG0310 and Young Elite Scientist Sponsorship Program (No. YESS20220409). The computations were performed on the High-Performance Computing (HPC) Platform of Hefei University of Technology.

References

An, D.; Qi, Y.; Li, Y.; Huang, Y.; Wang, L.; Tan, T.; and Shao, J. 2023. Bevbert: Multimodal map pre-training for language-guided navigation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2737–2748.

An, D.; Wang, H.; Wang, W.; Wang, Z.; Huang, Y.; He, K.; and Wang, L. 2024. Etpnav: Evolving topological planning for vision-language navigation in continuous environments. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Anderson, P.; Wu, Q.; Teney, D.; Bruce, J.; Johnson, M.; Sünderhauf, N.; Reid, I.; Gould, S.; and Van Den Hengel, A. 2018. Vision-and-language navigation: Interpreting visually-grounded navigation instructions in real environments. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 3674–3683.

Chang, A.; Dai, A.; Funkhouser, T.; Halber, M.; Niessner, M.; Savva, M.; Song, S.; Zeng, A.; and Zhang, Y. 2017. Matterport3D: Learning from RGB-D Data in Indoor Environments. *International Conference on 3D Vision (3DV)*.

Chen, J.; Lin, B.; Liu, X.; Ma, L.; Liang, X.; and Wong, K.-Y. K. 2025. Affordances-oriented planning using foundation models for continuous vision-language navigation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 23568–23576.

Chen, S.; Guhur, P.-L.; Schmid, C.; and Laptev, I. 2021. History aware multimodal transformer for vision-and-language navigation. *Advances in neural information processing systems*, 34: 5834–5847.

Chen, S.; Guhur, P.-L.; Tapaswi, M.; Schmid, C.; and Laptev, I. 2022a. Learning from unlabeled 3d environments for vision-and-language navigation. In *European Conference on Computer Vision*, 638–655. Springer.

Chen, S.; Guhur, P.-L.; Tapaswi, M.; Schmid, C.; and Laptev, I. 2022b. Think global, act local: Dual-scale graph transformer for vision-and-language navigation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 16537–16547.

Das, A.; Datta, S.; Gkioxari, G.; Lee, S.; Parikh, D.; and Batra, D. 2018. Embodied Question Answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; and Fei-Fei, L. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, 248–255. Ieee.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, 4171–4186.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. 2020. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *International Conference on Learning Representations*.

Fried, D.; Hu, R.; Cirik, V.; Rohrbach, A.; Andreas, J.; Morency, L.-P.; Berg-Kirkpatrick, T.; Saenko, K.; Klein, D.; and Darrell, T. 2018. Speaker-follower models for vision-and-language navigation. *Advances in Neural Information Processing Systems*, 31.

Gao, F.; Tang, J.; Wang, J.; Li, S.; and Yu, J. 2024. Enhancing Scene Understanding for Vision-andLanguage Navigation by Knowledge Awareness. *IEEE Robotics and Automation Letters*.

Georgakis, G.; Schmeckpeper, K.; Wanchoo, K.; Dan, S.; Miltsakaki, E.; Roth, D.; and Daniilidis, K. 2022. Cross-modal map learning for vision and language navigation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 15460–15470.

He, K.; Si, C.; Lu, Z.; Huang, Y.; Wang, L.; and Wang, X. 2024. Frequency-Enhanced Data Augmentation for Vision-and-Language Navigation. *Advances in Neural Information Processing Systems*, 36.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778.

Hong, Y.; Zhou, Y.; Zhang, R.; Dernoncourt, F.; Bui, T.; Gould, S.; and Tan, H. 2023. Learning navigational visual representations with semantic map supervision. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 3055–3067.

Hu, R.; Fried, D.; Rohrbach, A.; Klein, D.; Darrell, T.; and Saenko, K. 2019. Are You Looking? Grounding to Multiple Modalities in Vision-and-Language Navigation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 6551–6557.

Huang, C.; Mees, O.; Zeng, A.; and Burgard, W. 2023. Visual language maps for robot navigation. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, 10608–10615. IEEE.

Hwang, M.; Jeong, J.; Kim, M.; Oh, Y.; and Oh, S. 2023. Meta-explore: Exploratory hierarchical vision-and-language navigation using scene object spectrum grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6683–6693.

Ilinykh, N.; Emampoor, Y.; and Dobnik, S. 2022. Look and answer the question: On the role of vision in embodied question answering. In *Proceedings of the 15th International Conference on Natural Language Generation*, 236–245.

Krantz, J.; Wijmans, E.; Majundar, A.; Batra, D.; and Lee, S. 2020. Beyond the Nav-Graph: Vision and Language Navigation in Continuous Environments. In *European Conference on Computer Vision (ECCV)*.

Li, J.; and Bansal, M. 2023. Panogen: Text-conditioned panoramic environment generation for vision-and-language navigation. *Advances in Neural Information Processing Systems*, 36: 21878–21894.

Li, J.; Li, D.; Savarese, S.; and Hoi, S. 2023a. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, 19730–19742. PMLR.

Li, X.; Wang, Z.; Yang, J.; Wang, Y.; and Jiang, S. 2023b. Kerm: Knowledge enhanced reasoning for vision-and-language navigation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2583–2592.

Lin, B.; Nie, Y.; Wei, Z.; Zhu, Y.; Xu, H.; Ma, S.; Liu, J.; and Liang, X. 2024. Correctable Landmark Discovery Via Large Models for Vision-Language Navigation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Lin, B.; Zhu, Y.; Liang, X.; Lin, L.; and Liu, J. 2023. Actional atomic-concept learning for demystifying vision-language navigation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, 1568–1576.

Liu, R.; Wang, W.; and Yang, Y. 2024. Volumetric environment representation for vision-language navigation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 16317–16328.

Pan, B.; Panda, R.; Jin, S.; Feris, R.; Oliva, A.; Isola, P.; and Kim, Y. 2024. LangNav: Language as a Perceptual Representation for Navigation. In *Findings of the Association for Computational Linguistics: NAACL 2024*, 950–974.

Pashevich, A.; Schmid, C.; and Sun, C. 2021. Episodic transformer for vision-and-language navigation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 15942–15952.

Qi, Y.; Wu, Q.; Anderson, P.; Wang, X.; Wang, W. Y.; Shen, C.; and Hengel, A. v. d. 2020. Reverie: Remote embodied visual referring expression in real indoor environments. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9982–9991.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763. PMLR.

Savva, M.; Kadian, A.; Maksymets, O.; Zhao, Y.; Wijmans, E.; Jain, B.; Straub, J.; Liu, J.; Koltun, V.; Malik, J.; et al. 2019. Habitat: A platform for embodied ai research. In *Proceedings of the IEEE/CVF international conference on computer vision*, 9339–9347.

Sigurdsson, G. A.; Thomason, J.; Sukhatme, G. S.; and Piramuthu, R. 2023. Rrex-bot: Remote referring expressions with a bag of tricks. In *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 5203–5210. IEEE.

Tan, S.; Ge, M.; Guo, D.; Liu, H.; and Sun, F. 2022. Depth-aware vision-and-language navigation using scene query attention network. In *2022 International Conference on Robotics and Automation (ICRA)*, 9390–9396. IEEE.

Wang, L.; He, Z.; Dang, R.; Shen, M.; Liu, C.; and Chen, Q. 2024a. Vision-and-Language Navigation via Causal Learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 13139–13150.

Wang, L.; He, Z.; Tang, J.; Dang, R.; Wang, N.; Liu, C.; and Chen, q. 2023a. A Dual Semantic-Aware Recurrent Global-Adaptive Network For Vision-and-Language Navigation. In *IJCAI*.

Wang, L.; Tang, J.; He, Z.; Dang, R.; Liu, C.; and Chen, Q. 2024b. Enhanced Language-guided Robot Navigation with Panoramic Semantic Depth Perception and Cross-modal Fusion. In *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 7726–7733. IEEE.

Wang, X.; Huang, Q.; Celikyilmaz, A.; Gao, J.; Shen, D.; Wang, Y.-F.; Wang, W. Y.; and Zhang, L. 2019. Reinforced cross-modal matching and self-supervised imitation learning for vision-language navigation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 6629–6638.

Wang, Z.; Li, J.; Hong, Y.; Wang, Y.; Wu, Q.; Bansal, M.; Gould, S.; Tan, H.; and Qiao, Y. 2023b. Scaling data generation in vision-and-language navigation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 12009–12020.

Wang, Z.; Li, X.; Yang, J.; Liu, Y.; and Jiang, S. 2023c. Gridmm: Grid memory map for vision-and-language navigation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 15625–15636.

Wu, J.; Wu, C.; Shen, X.; and Wang, L. 2025. Adaptive Cross-Modal Experts Network with Uncertainty-Driven Fusion for Vision–Language Navigation. *Knowledge-Based Systems*, 307: 112735.

Xia, F.; Zamir, A. R.; He, Z.; Sax, A.; Malik, J.; and Savarese, S. 2018. Gibson env: Real-world perception for embodied agents. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 9068–9079.

Zhan, Z.; Qin, J.; Zhuo, W.; and Tan, G. 2024. Enhancing Vision and Language Navigation with Prompt-based Scene Knowledge. *IEEE Transactions on Circuits and Systems for Video Technology*.

Zhang, X.; He, J.; Zhao, J.; Hu, Z.; Yang, X.; Li, J.; and Hong, R. 2024. Exploring and exploiting model uncertainty for robust visual question answering. *Multimedia Systems*, 30(6): 1–14.

Zhu, F.; Liang, X.; Zhu, Y.; Yu, Q.; Chang, X.; and Liang, X. 2021. Soon: Scenario oriented object navigation with graph-based exploration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 12689–12699.