A Dual Semantic-Aware Recurrent Global-Adaptive Network for Vision-and-Language Navigation

Liuyi Wang[1] , **Zongtao He**[1] , **Jiagui Tang**[1] , **Ronghao Dang**[1] , **Naijia Wang**[1] , **Chengju Liu**[1,2*] and **Qijun Chen**[1*]
[1]Tongji University, Shanghai, China
[2]Tongji Artificial Intelligence (Suzhou) Research Institute, Suzhou, China
{wly, xingchen327, 2130701, dangronghao, 2030715, liuchengju, qjchen}@tongji.edu.cn

Abstract

Vision-and-Language Navigation (VLN) is a realistic but challenging task that requires an agent to locate the target region using verbal and visual cues. While significant advancements have been achieved recently, there are still two broad limitations: (1) The explicit information mining for significant guiding semantics concealed in both vision and language is still under-explored; (2) The previously structured map method provides the average historical appearance of visited nodes, while it ignores distinctive contributions of various images and potent information retention in the reasoning process. This work proposes a dual semantic-aware recurrent global-adaptive network (DSRG) to address the above problems. First, DSRG proposes an instruction-guidance linguistic module (IGL) and an appearance-semantics visual module (ASV) for boosting vision and language semantic learning respectively. For the memory mechanism, a global adaptive aggregation module (GAA) is devised for explicit panoramic observation fusion, and a recurrent memory fusion module (RMF) is introduced to supply implicit temporal hidden states. Extensive experimental results on the R2R and REVERIE datasets demonstrate that our method achieves better performance than existing methods. Code is available at https://github.com/CrystalSixone/DSRG.

1 Introduction

As an important application in human-robot interaction, the vision-language navigation (VLN) task [Anderson *et al.*, 2018] has attracted considerable attention. It is a crucial but challenging task that requires an embodied agent to reach the required locations in unstructured environments only based on visual observations and given verbal instructions. The main issue relies on how to effectively mine and exploit the high-level context connotation hidden in the plentiful features of vision and language so that they can serve as better guiding elements in the serial navigation process.

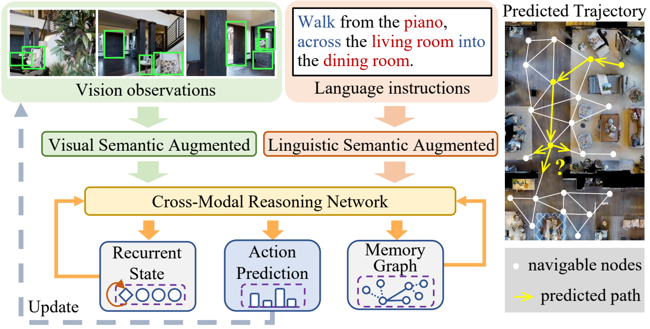

Vision observations
Walk from the piano, across the living room into the dining room.
Language instructions
Predicted Trajectory
Visual Semantic Augmented
Linguistic Semantic Augmented
Cross-Modal Reasoning Network
Recurrent State
Action Prediction
Memory Graph
Update
?
navigable nodes
predicted path

Figure 1: Illustration of the proposed DSRG, which first augments semantics of visual and linguistic inputs respectively, and employs the cross-modal reasoning network along with the recurrent state fusion module and the memory graph to predict the action.

Due to the huge success of the transformer [Vaswani *et al.*, 2017], transformer-based cross-modal fusion methods [Hao *et al.*, 2020; Hong *et al.*, 2021; Zhu and Yang, 2020; Chen *et al.*, 2022b] have undergone substantial development and show promise. However, these methods still have some limitations. First, existing cross-modal networks don't leverage the guiding semantic information hidden in different inputs sufficiently. Intuitively, the guiding semantic features are both hidden in visual inputs (e.g., the object features in images) and linguistic inputs (e.g., the direction and landmark tokens). However, some recent semantic-related approaches [Qi *et al.*, 2021; Zhang *et al.*, 2021; An *et al.*, 2021] either focus on the semantic enhancement of only one modality or employ several independent soft attention modules to learn correlation in special embeddings, without considering the simultaneous explicit modeling of finer visual and linguistic semantics for precise perception.

Secondly, the history-dependent ability of current models to infer the action for each step is inadequate. The long-term decision-making process requires the agent to keep track of the exploration progress with respect to its corresponding sub-instruction. While some recent transformer-based methods [Chen *et al.*, 2021b; Chen *et al.*, 2022c; Chen *et al.*, 2022a] structurally model the historical information of visited nodes as the graph map, they simply provide average exterior features of past observations, lacking the flow of reasoning presentation within the network, which

[*]corresponding author

could lead to information loss and reasoning discontinuity.

To address the above problems, we propose DSRG, a novel dual semantic-aware recurrent global-adaptive network for VLN, where the guiding semantic features concealed in the visual and linguistic representations are better utilized to guide the navigation, and the memory augmentation is implemented based on the explicit outer flow (the structured graph of visited nodes) and the implicit inner flow (the specific recurrent memory token). To achieve these two goals, as shown in Fig. 1, we first design a dual structure, including an instruction-guidance linguistic module (IGL) and an appearance-semantics visual module (ASV), to learn guiding semantic presentations of two types of inputs, respectively. Based on the dominant semantic prior, the agent is capable to capture the key components of visual and verbal inputs. Then, for the memory mechanism, we first adopt the global map used in [Chen *et al.*, 2022c] with a proposed global adaptive aggregation method (GAA) to fuse panoramic observations of visited nodes by adaptively learning the contribution of various views to candidate sites. Further, to improve the continuity of reasoning procedures within the network, we suggest a recurrent memory fusion module (RMF) to broadcast the inference states in a recurrent way. Experimental results on two datasets, R2R [Anderson *et al.*, 2018] and REVERIE [Qi *et al.*, 2020b], have proved the effectiveness of our method, achieving a new state-of-the-art borderline on the VLN task.

Overall, our contributions are summarised as follows: (1) We propose a dual semantic-augmented structure to boost visual and linguistic semantic representations, respectively. (2) We propose to use both explicit and implicit memory transfer channels for enhancing the model's ability to adaptively memorize and infer the status of navigation. (3) Extensive experiments on two datasets, R2R and REVERIE, demonstrate that the proposed DSRG outperforms other existing methods.

2 Related Work

The vision-and-language navigation (VLN) task is first proposed by [Anderson *et al.*, 2018], which serves as a new technique for relating natural language to vision and action in unstructured and unseen environments. To improve the generalization of the agent in unseen environments, some environment-augmented methods [Fried *et al.*, 2018; Tan *et al.*, 2019; Liu *et al.*, 2021; Li *et al.*, 2022; Liang *et al.*, 2022b; Wang *et al.*, 2023] are proposed for improving the diversity of scenes and instructions. Additionally, the task-specific auxiliary task methods [Ma *et al.*, 2019; Zhu *et al.*, 2020; Zhao *et al.*, 2022] are proposed to enhance the interpretability and navigation ability of the model. In general, most VLN methods encode visual and linguistic features via the large pre-trained networks, without considering the explicit usage of the semantic-level cues of the two types of inputs which are crucial for directing the agent.

Semantics in VLN. Recently, some methods have demonstrated the advantages of semantic features for vision-based navigation [Dang *et al.*, 2022a; Dang *et al.*, 2022b]. ORIST [Qi *et al.*, 2021] and SOAT [Moudgil *et al.*, 2021] propose to concatenate object-level features with the scene-level features and learn them through the transformer encoder in a parallel way. BiasVLN [Zhang *et al.*, 2021] observes that the low-level image features result in environmental bias. SEvol [Chen *et al.*, 2022a] utilizes a graph-based method to construct relationships of objects. However, all of the above methods only boost semantics from the visual perspective and neglect guiding hints hidden in natural language instructions. OAAM [Qi *et al.*, 2020a] and NvEM [An *et al.*, 2021] apply independent soft attention to text embeddings to learn the object and action representations of instructions, and the latter also considers the neighboring objects of candidates. ADAPT [Lin *et al.*, 2022] suggests using the CLIP [Radford *et al.*, 2021] with an action prompt to improve the action-level modality alignment. Different from earlier methods that only focus on the semantics of a single modality or implicitly learn semantics based on soft attention with hidden states, we argue that the crucial guiding information exists in both vision and language and should be explicitly highlighted. Therefore, we propose a dual structure to enhance semantic features for vision and language presentations by injecting the extracted prior knowledge and then fusing them through the adaptively global-local cross-modal module.

Historical Memory in VLN. It is crucial for agents to represent both the current and prior states while navigating. For LSTM-based methods [Wang *et al.*, 2020; An *et al.*, 2021; Wang *et al.*, 2021], the long-term historical information is broadcast via the inherent hidden state of the network. With the enormous success made by the transformer [Vaswani *et al.*, 2017], a growing number of models have lately achieved better performance based on the transformer structure. Some pre-trained models [Hao *et al.*, 2020; Guhur *et al.*, 2021; Liang *et al.*, 2022a] are proposed to improve the model's capability of representation via the synthesized dataset. To provide the model with historical observations and improve inference capability, some methods [Chen *et al.*, 2021a; Chen *et al.*, 2021b; Chen *et al.*, 2022c] focus on constructing the structured memory graph of visited nodes. However, these methods neglect to transmit the network's reasoning states at each step, which might lead to the interruption of the inference process. Inspired by RecBERT [Hong *et al.*, 2021], we propose to improve the global memory graph adopted in the previous SOTA method [Chen *et al.*, 2022c] by a global adaptive aggregation method and then implement a recurrent fusion module to present the current reasoning states.

3 Methodology

Based on the given instruction, the task of VLN is to predict sequence actions toward the target location automatically. Our goal is to take full advantage of the semantic features hidden in linguistic and visual representations to achieve more accurate navigation. As shown in Fig. 2, the overall framework of our proposed method can be divided into three sub-modules: (a) guidance-aware linguistic instruction learning (Sec. 3.1), (b) semantic-aware visual environment learning (Sec. 3.2) and (c) recurrent global-local visual-linguistic feature fusion (Sec. 3.3). Specifically, we first use a dual structure to enrich vision and language expressions by injecting guiding semantic features in their respective domains. Next, we construct a global navigation map via the adaptive

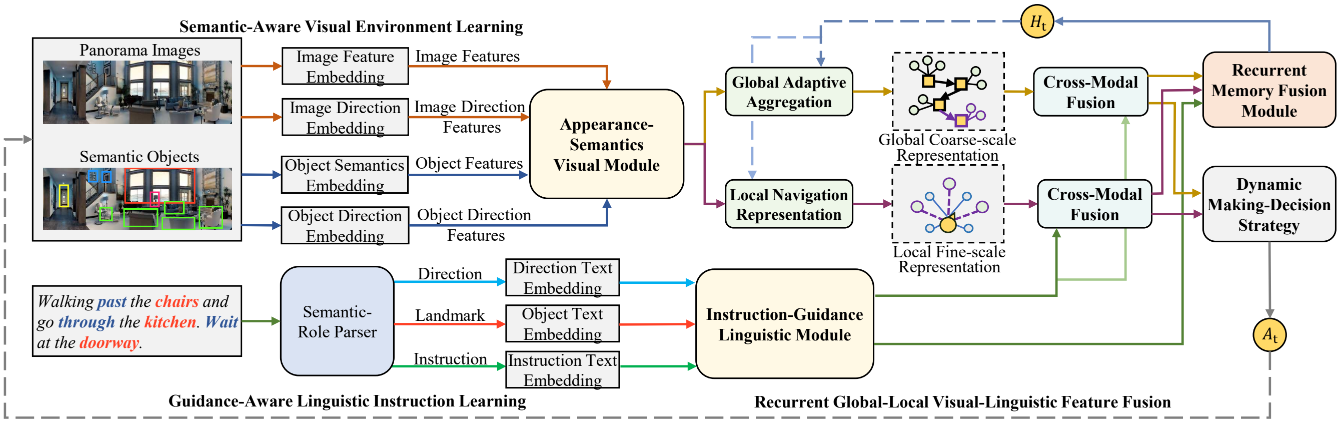

Semantic-Aware Visual Environment Learning
Panorama Images
Semantic Objects
Image Feature Embedding
Image Direction Embedding
Object Semantics Embedding
Object Direction Embedding
Image Features
Image Direction Features
Object Features
Object Direction Features
Appearance-Semantics Visual Module
Global Adaptive Aggregation
Local Navigation Representation
Global Coarse-scale Representation
Local Fine-scale Representation
Cross-Modal Fusion
Cross-Modal Fusion
Recurrent Memory Fusion Module
Dynamic Making-Decision Strategy
H_t
A_t
Walking past the chairs and go through the kitchen. Wait at the doorway.
Semantic-Role Parser
Direction
Landmark
Instruction
Direction Text Embedding
Object Text Embedding
Instruction Text Embedding
Instruction-Guidance Linguistic Module
Guidance-Aware Linguistic Instruction Learning
Recurrent Global-Local Visual-Linguistic Feature Fusion

Figure 2: Overview of the proposed DSRG structure, which includes three components: (a) semantic-aware visual environment learning, (b) guidance-aware linguistic instruction learning, and (c) recurrent global-local visual-linguistic feature fusion.

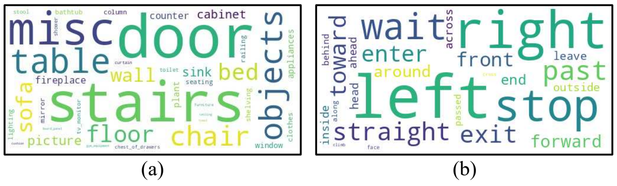


Figure 3: Word clouds of (a) landmark and (b) direction tokens.

fusion module to assist the local decision-making process. The inherent memory unit is updated based on the instruction, global map and local observations, and transferred to the next action step. Last, the possibility of action is calculated by the dynamic decision-making strategy.

3.1 Guidance-Aware Linguistic Instruction Learning

The natural language contains a wealth of semantic information, serving as a vital link for social interaction. In this work, we first extract guiding semantic phrases from instructions and then encourage the model to focus on these dominant parts and enhance the instruction representations.

Semantic-Role Parser. As shown in Fig. 3, two kinds of semantic information are specified: direction-level and landmark-level phrases, where the former can guide the agent's forward direction and the latter can provide the referential landmark during navigation. While it is easy for humans to quickly capture these leading phrases, the network can only gradually learn this from massive data fitting. Therefore, we extract them based on the toolkit NLTK [Bird *et al.*, 2009] to perform direction-landmark recognition based on its part-of-speech tags and entity dictionary. Following the category map provided by MP3D [Chang *et al.*, 2017], all extracted entities are normalized into 43 categories.

Instruction-Guidance Linguistic Module (IGL). As shown in Fig. 4, for the whole instruction $I = \{w_1, ..., w_L\}$, direction-level phrases $I_d = \{w_1^d, ..., w_{L_d}^d\}$, and landmark-level phrases $I_e = \{w_1^e, ..., w_{L_e}^e\}$, each token is firstly respectively embedded into a vector of 768-dimension.

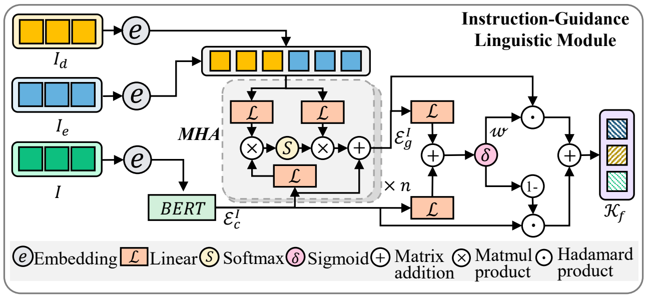


Figure 4: Illustration of the instruction-guidance linguistic module.

The position embedding function is employed to inject the ordering relation. Then the tokens in I are encoded by a BERT [Devlin *et al.*, 2019] model to get context linguistic features denoted as $\mathcal{E}_c^I$. To highlight the guided representations of the key areas in the instruction as $\mathcal{E}_g^I$, we propose to use the multi-head attention (MHA) mechanism [Vaswani *et al.*, 2017] to update the context features based on the correlation between context tokens and the concatenate of direction and landmark tokens. Then, to balance the proportion of context content $\mathcal{E}_c^I$ and guided semantic features $\mathcal{E}_g^I$, we use a gate-like structure with a sigmoid function $\delta(\cdot)$ to dynamically obtain the weight value $\omega \in \mathbb{R}^{L\times 1}$ and then weighted sum the above two kinds of features into $\mathcal{K}_f \in \mathbb{R}^{L\times d_m}$. The above process is formulated as Eq. (1) – (5):

$$(\mathcal{I}, \mathcal{I}_d, \mathcal{I}_e) = \text{Embedding}(I, I_d, I_e) \tag{1}$$

$$\mathcal{E}_c^I = \text{BERT}(\mathcal{I}) \tag{2}$$

$$\mathcal{E}_g^I = \text{MHA}(\mathcal{E}_c^I, [\mathcal{I}_d, \mathcal{I}_e]) \tag{3}$$

$$\omega = \delta(\mathcal{E}_g^I W_g + \mathcal{E}_c^I W_c + b_I) \tag{4}$$

$$\mathcal{K}_f = \omega \odot \mathcal{E}_g^I + (1-\omega) \odot \mathcal{E}_c^I \tag{5}$$

where $[.,.]$ denotes concatenation operation, $W_g \in \mathbb{R}^{d_m\times 1}$ and $W_c \in \mathbb{R}^{d_m\times 1}$ present the learnable parameters.

3.2 Semantic-Aware Visual Environment Learning

In VLN, another important modality of input is visual observation as the agent can only explore the unstructured environments based on images. Therefore, we further devise a semantic-aware visual environment learning approach to form a dual semantic-aware structure with the enhanced language features discussed in Sec. 3.1. By considering the fine-grained semantic object features within each sub-image of the panorama and fusing them with the image features, the environmental presentations obtain semantic enhancement.

Visual Image Feature. The connected navigation graphs are specified by the Matterport3D simulator [Chang *et al.*, 2017], where the navigable nodes are given discretely. Formally, each panorama is split into 36 images $V = \{v_i\}_{i=1}^{36}$. The corresponding heading θ and elevation γ are denoted as $V_r = \{r_i\}_{i=1}^{36}$, where $r_i = (\sin\theta_i, \cos\theta_i, \sin\gamma_i, \cos\gamma_i) \in \mathbb{R}^4$. The ViT [Dosovitskiy *et al.*, 2020] model is employed to extract image features. Aggregated by the angle features $V_r \in \mathbb{R}^{36\times4}$, token types $V_t \in \mathbb{R}^{36\times1}$ (initialized as ones), and navigable types $V_n \in \mathbb{R}^{36\times1}$ (one for the navigable and zero for the non-navigable), the visual image features $\mathcal{E}_f^V \in \mathbb{R}^{36\times d_m}$ are formulated as Eq. (6) – (8):

$$[\mathcal{E}_r^V, \mathcal{E}_t^V, \mathcal{E}_n^V] = [V_r, V_t, V_n]W_z + b_z \tag{6}$$

$$\mathcal{E}_v^V = \mathrm{ViT}(V)W_v \tag{7}$$

$$\mathcal{E}_f^V = \mathrm{LN}(\mathcal{E}_v^V + \mathcal{E}_r^V + \mathcal{E}_t^V + \mathcal{E}_n^V) \tag{8}$$

where $W_z \in \mathbb{R}^{6\times d_m}, b_z \in \mathbb{R}^{d_m}, W_v \in \mathbb{R}^{d_v\times d_m}$.

Semantic Object Feature. The previous methods [Moudgil *et al.*, 2021; Qi *et al.*, 2021] simply concatenate objects and images in parallel without considering respective object features in 36 images of V, which may lead to the problem of feature dislocation. In contrast, we propose to use the fine-grained object features of 36 images for each viewpoint, which can be denoted as $O = \{[o_{i1}, ..., o_{iM}]\}_{i=1}^{36}$, where M is the maximum number of objects in each image. Let $O_g \in \mathbb{R}^{36\times M\times3}, O_r \in \mathbb{R}^{36\times M\times4}, O_l \in \mathbb{R}^{36\times M\times1}, O_n \in \mathbb{R}^{36\times M\times1}$ denote the geometric features, angle features, label features, and navigable types, respectively. To better leverage the semantic information, we filter out some less informative items (e.g., walls and ceilings) and take the first M items with the largest area. The semantic object features $\mathcal{E}_f^O \in \mathbb{R}^{36\times M\times d_o}$ are computed as follows:

$$[\mathcal{E}_g^O, \mathcal{E}_r^O, \mathcal{E}_l^O, \mathcal{E}_n^O] = [O_g, O_r, O_l, O_n]W_x + b_x \tag{9}$$

$$\mathcal{E}_f^O = \mathrm{LN}(\mathcal{E}_g^O + \mathcal{E}_r^O + \mathcal{E}_l^O + \mathcal{E}_n^O) \tag{10}$$

where $W_x \in \mathbb{R}^{9\times d_o}$ and $b_x \in \mathbb{R}^{d_o}$ are learnable parameters.

Appearance-Semantics Visual Module (ASV). As there is a large overlap space between 36 images, the objects have natural internal relationships with each other as well. Therefore, as shown in Fig. 5, we first reshape object features from different images as $\widetilde{\mathcal{E}}_f^O \in \mathbb{R}^{(36\times M)\times d_o}$ and use a transformer block to promote intermediate semantic-level representations. Similarly, another transformer is used to learn the relationships between images, given by

$$\widetilde{\mathcal{S}}_f^O = \mathrm{Trans}_o(\widetilde{\mathcal{E}}_f^O),\ \mathcal{S}_f^V = \mathrm{Trans}_v(\mathcal{E}_f^V) \tag{11}$$

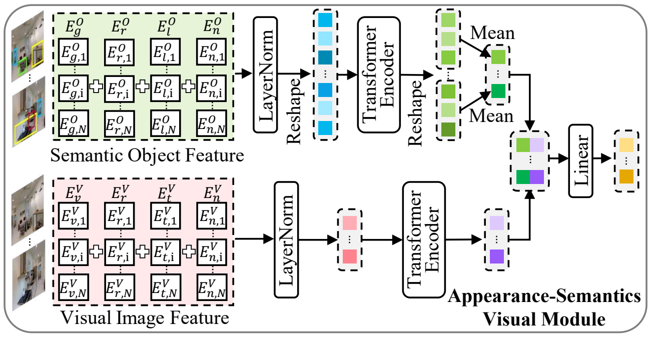


Figure 5: Illustration of the appearance-semantics visual module.

where Trans_o and Trans_v mean the standard transformer block consisting of multi-head self-attention, residual connection, layernorm (LN) and feed-forward network (FFN). After obtaining the cross-image object feature $\widetilde{\mathcal{S}}_l^O$, we reshape it along the image dimension and then calculate the average representation of objects in each image. Finally, we concatenate the appearance-level features $\mathcal{S}_f^V$ and semantics-level features $\mathcal{S}_l^O$ together to obtain the semantic-aware visual representations as:

$$\mathcal{S}_f^O = \frac{1}{M}\sum_{i=1}^{M}\widetilde{\mathcal{S}}_{f,i}^O \tag{12}$$

$$\mathcal{S}_f = \mathrm{LN}([\mathcal{S}_f^V, \mathcal{S}_l^O]W_f + b_f) \tag{13}$$

where $W_f \in \mathbb{R}^{(d_m+d_o)\times d_m}, b_f \in \mathbb{R}^{d_m}$ are learnable parameters. In this work we employ $d_o = 128$.

3.3 Recurrent Global-Local Visual-Linguistic Feature Fusion

The previous method [Chen *et al.*, 2022c] has shown the effectiveness of using a global memory graph based on average visual observations to promote inference capability. We argue that this is necessary but insufficient since it ignores the different contributions of images in panorama, and the reasoning states within the network have not got fully transferred through the navigation steps. Therefore, in this section, we devise a global adaptive aggregation method and a recurrent memory fusion module for memory augmentation.

Local Navigation Representation. To provide fine-scale observation of the current node, the representations in the local branch are first comprised of enhanced semantic-aware visual features $\mathcal{S}_f$ described in Sec. 3.2, concatenated by a zero-initialized `[CLS]` token (which also serves as the stop token) and the recurrent memory token `[MEM]`. The structure of the local navigation representation can be formulated as $\mathcal{U}_L = (\texttt{[CLS]}, \mathcal{S}_f, \texttt{[MEM]})$, where the calculation of `[MEM]` will be explained in the following. The embedding of the local position features is further added to $\mathcal{U}_l$ as $\widetilde{\mathcal{U}}_l$ to present the relative position information.

Global Adaptive Aggregation (GAA). In DUET [Chen *et al.*, 2022c], the global navigation map is constructed by storing visited nodes, navigable nodes, and the current node with their appearance features at each step, which has shown superior performance. Specifically, the visual representation of each node is represented by the average function $\frac{1}{N}\sum_{i=1}^{N}\mathcal{E}_{f,i}^{V}$, where N denotes the length of tokens in each panoramic observation. However, this operation flattens the different contributions of different images. Intuitively, the images that do not contain specific landmarks or are far aware from candidate points have smaller contributions. It is necessary to pay more attention to more instruction-relevant parts to guide navigation. Therefore, we propose the GAA module to make the network adaptively learn to encode relationships between images. Supposed $\mathcal{S}_f = \{s_i\}_{i=1}^{N} \in \mathbb{R}^{N\times d_m}$, we introduce an attention matrix $W_f \in \mathbb{R}^{d_m\times 1}$ with the softmax function to adaptively re-weight image features and sum them up, achieving the aggregated features $\mathcal{S}_a \in \mathbb{R}^{1\times d_m}$:

$$\mathcal{R} = [r_1, ..., r_N] = \gamma(\mathcal{S}_f W_f + b_f) \tag{14}$$

$$\widetilde{r}_i = \frac{e^{r_i}}{\sum_{j\in N} e^{r_j}} \tag{15}$$

$$\mathcal{S}_a = \sum_{i\in N} \widetilde{r}_i * s_i \tag{16}$$

where $\gamma(\cdot)$ is the tanh activation function. Let $\mathcal{S}_g$ denote the set of aggregated node features in the global map, similar to the local branch, the structure of the global map representation is formulated as $\mathcal{U}_g = (\texttt{[CLS]}, \mathcal{S}_g, \texttt{[MEM]})$. The embedding of step position is also added to $\mathcal{U}_g$ as $\widetilde{\mathcal{U}}_g$.

Cross-Modal Feature Fusion. After obtaining the semantic-augmented features for both vision and language, it is essential to fuse these features together to learn the correlation between these two kinds of inputs. LXMERT [Tan and Bansal, 2019] is adopted as the cross-modal encoder. Following [Hong *et al.*, 2021], the language tokens are only assigned as keys and values to update the visual tokens during the fine-tuning. Concretely, two independent cross-modal fusion modules are employed to fuse the linguistic features $\mathcal{K}_f$ with the global and local visual features $\{\widetilde{\mathcal{U}}_g, \widetilde{\mathcal{U}}_l\}$, achieving $\mathcal{F}_g$ and $\mathcal{F}_l$, respectively.

Recurrent Memory Fusion Module (RMF). Considering the construction of the global map is an explicit memory representation, we propose to further build an RMF module to transmit the network's intermediate reasoning states. Intuitively, in the sequential reasoning task, we can always remember the basis of our own judgment at the last moment, which can make the subsequent reasoning easier and more accurate. Therefore, for the t-th step, we extract the [CLS] tokens from $\mathcal{F}_g$, $\mathcal{F}_l$ and $\mathcal{K}_f$ as $\mathcal{C}_g, \mathcal{C}_l$ and $\mathcal{C}_k$, respectively. This is because [CLS] tokens can be regarded as the highly fused presentations of the corresponding modalities at the current step. Then we project them to the memory representation domains via the linear transformations after concatenating:

$$\mathcal{H}_r = \mathrm{LN}([\mathcal{C}_g, \mathcal{C}_l, \mathcal{C}_k]W_c + b_c) \tag{17}$$

where $W_c \in \mathbb{R}^{3d_m\times d_m}$ and $b \in \mathbb{R}^{d_m}$. To avoid interference in the prediction of the stop signal, a separate token [MEM] is defined to specifically store the hidden reasoning representation in each step. The obtained inherent memory unit $\mathcal{H}_r \in \mathbb{R}^{1\times d_m}$ is assigned by the recurrent token [MEM] of the global and local sequence at the next step.

Dynamic Decision-Making Strategy. Finally, we employ the dynamic decision-making strategy proposed by [Chen *et al.*, 2022c] to predict the action. Concretely, the one-dimensional scalar weight σ of global and local branches are computed by the FFN network on the concatenation of $\mathcal{C}_g$ and $\mathcal{C}_l$, respectively. After using another two FFN networks to project global-local fused features into the score domains, the local action scores $\mathcal{G}_l$ are converted into the global action space $\hat{\mathcal{G}}_l$. Then the final probability of the action prediction $\mathcal{G}$ is obtained via the weighted sum of the two branches. Only candidate nodes will be considered based on the mask function. The formulas are as Eq. (18) – (20):

$$\sigma = \delta(\mathrm{FFN}([\mathcal{C}_g, \mathcal{C}_l])) \tag{18}$$

$$\mathcal{G}_g = \mathrm{FFN}(\mathcal{F}_g),\ \mathcal{G}_l = \mathrm{FFN}(\mathcal{F}_l) \tag{19}$$

$$\mathcal{G} = \sigma\mathcal{G}_g + (1-\sigma)\hat{\mathcal{G}}_l \tag{20}$$

The cross-entropy loss is used as the optimization objective:

$$\mathcal{L} = \sum_{t=1}^{T} -\log p(a_t^*|\mathcal{I}, \mathcal{V}_t, a_{1:t-1}) \tag{21}$$

4 Experiments

4.1 Dataset and Implementation Details

Dataset. To validate our proposed method, we conduct extensive experiments on the R2R [Anderson *et al.*, 2018] and REVERIE datasets [Qi *et al.*, 2020b]. Based on 90 different buildings, R2R includes 21,576 fine-grained step-by-step instructions to guide the agent, and REVERIE includes 21,702 shorter annotated instructions for remoted referring expressions. The fine-grained object features are provided by REVERIE, where each panoramic sub-image contains an average of 10 objects.

Evaluation Metrics. For R2R, four standard metrics are for evaluation: the navigation error (NE): the distance between the ground truth and the agent's stop position; the success rate (SR): the ratio of paths that stop within 3m from the target points; the oracle success rate (OSR): SR with the oracle stop policy; and the success rate weighted by the path length (SPL): SR penalized by the path length. For REVERIE, another two metrics are added: remote grounding success rate (RGS): the ratio of objects grounded correctly, and the RGS weighted by the path length (RGSPL).

Implementation Details. In the pre-training stage, we train our DSRG with batch size 24 for 400k iterations using 1 NVIDIA RTX 3090 GPU. Since the supplement of directional and semantic texts will cause the leakage for the masked language modeling (MLM) [Devlin *et al.*, 2019], and the recurrent state cannot be directly used for the single-step action prediction (SAP) [Chen *et al.*, 2021b], we did not enable IGL and RMF in the pre-training phase, but added them in the fine-tuning stage. During fine-tuning, the batch size and the learning rate are 4 and 5×10^{-6}, respectively. We use the ViT-B/16 [Dosovitskiy *et al.*, 2020] pre-trained model to extract image features. The numbers of transformer layers for

Methods	Validation Seen				Validation Unseen				Test Unseen			
	NE↓	OSR↑	SR↑	SPL↑	NE↓	OSR↑	SR↑	SPL↑	NE↓	OSR↑	SR↑	SPL↑
EnvDrop [Tan *et al.*, 2019]	3.99	-	62	59	5.22	-	52	48	5.23	59	51	47
AuxRN [Zhu *et al.*, 2020]	3.33	78	70	67	5.28	62	55	50	5.15	62	55	51
PREVALENT [Hao *et al.*, 2020]	3.67	-	60	65	4.73	-	57	53	4.75	61	54	51
RecBERT [Hong *et al.*, 2021]	2.90	79	72	68	3.93	69	63	57	4.09	70	63	57
NvEM [An *et al.*, 2021]	3.44	-	69	65	4.27	-	60	55	4.37	66	58	54
HOP [Qiao *et al.*, 2022]	2.72	-	75	70	3.80	-	64	57	3.83	-	64	59
EnvMix [Liu *et al.*, 2021]	2.48	-	75	72	3.89	-	64	58	3.87	72	65	59
HAMT [Chen *et al.*, 2021b]	2.51	82	76	72	**2.29**	73	66	61	3.93	72	65	60
TD-STP [Zhao *et al.*, 2022]	2.34	83	77	73	3.22	76	70	**63**	3.73	72	67	61
DUET [Chen *et al.*, 2022c]	2.28	86	79	73	3.31	81	72	60	3.65	76	69	59
DSRG (Ours)	**2.23**	**88**	**81**	**76**	3.00	**81**	**73**	62	**3.33**	**78**	**72**	**61**

Table 1: Comparison with the state-of-the-art methods on the R2R dataset.

Methods	Val Seen					Val Unseen					Test Unseen				
	Navigation			Grounding		Navigation			Grounding		Navigation			Grounding	
	OSR↑	SR↑	SPL↑	RGS↑	RGSPL↑	OSR↑	SR↑	SPL↑	RGS↑	RGSPL↑	OSR↑	SR↑	SPL↑	RGS↑	RGSPL↑
RCM [Wang *et al.*, 2019]	29.44	23.33	21.82	16.23	15.36	14.23	9.29	6.97	4.89	3.89	11.68	7.84	6.67	3.67	3.14
MATTN [Qi *et al.*, 2020b]	55.17	50.53	45.50	31.97	29.66	28.20	14.40	7.19	7.84	4.67	30.63	19.88	11.61	11.28	6.08
SIA [Lin *et al.*, 2021]	65.85	61.91	57.08	45.96	42.65	44.67	31.53	16.28	22.41	11.56	44.56	30.80	14.85	19.02	9.20
RecBERT [Hong *et al.*, 2021]	53.90	41.79	47.96	38.23	35.61	35.02	30.67	24.90	18.77	15.27	32.91	29.61	23.99	16.50	13.51
HOP [Qiao *et al.*, 2022]	54.88	53.76	47.19	38.65	33.85	36.24	31.78	26.11	18.85	15.73	33.06	24.34	16.38	17.69	14.34
HAMT [Chen *et al.*, 2021b]	47.65	43.29	40.19	27.20	25.18	36.84	32.95	30.20	18.92	17.28	33.41	30.40	26.67	14.88	13.08
DUET [Chen *et al.*, 2022c]	73.86	71.75	63.94	57.41	51.14	51.07	46.98	33.73	32.15	23.03	56.91	52.51	36.06	31.88	22.06
DSRG (Ours)	**77.72**	**75.69**	**68.09**	**61.07**	**54.72**	**53.25**	**47.83**	**34.02**	**32.69**	**23.37**	**58.26**	**54.04**	**37.09**	**32.49**	**22.18**

Table 2: Comparison with the state-of-the-art methods on the REVERIE dataset.

Id	ASV	IGL	SR↑	SPL↑	NE↓	OSR↑
1	✗	✗	69.18	60.28	3.36	77.31
2	✓	✗	71.43	61.07	3.37	78.37
3	✗	✓	71.26	61.20	3.26	78.93
4	✓	✓	**72.50**	**61.56**	**3.00**	**80.97**

Table 3: Ablation study for the dual semantic-aware modules.

instructions, visual and semantic features, and local-global cross-modal attention modules are 9, 2 and 4, respectively.

4.2 Comparison with State-of-the-Arts

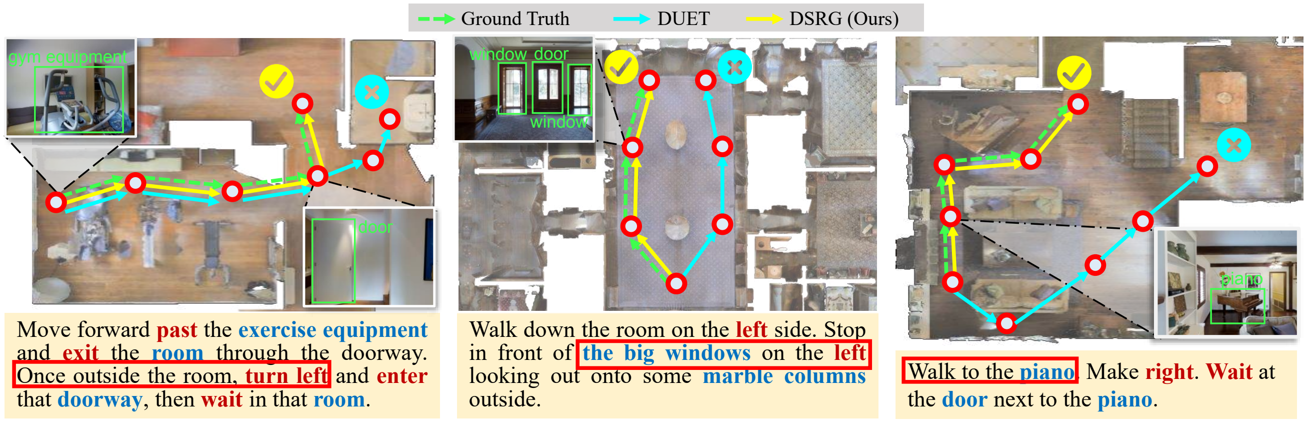
The quantitative performance results in comparison with state-of-the-art methods on R2R and REVERIE datasets are listed in Table 1 and Table 2, respectively. The proposed DSRG achieves the leading performance on both two datasets. Specifically, our model outperforms the current SOTA [Chen *et al.*, 2022c] significantly, with SPL being improved by 2% and 2% on the R2R unseen validation and test splits, respectively. The SR on REVERIE also obtains large improvements by about 4% and 3.5% on the seen and unseen splits, respectively. These results demonstrate that our model is beneficial for enhancing the performance of the agent in VLN, with a more fine-grained perception of the environmental and linguistic semantics and better long-term exploration understanding.

4.3 Ablation Study

We perform ablation studies to evaluate the proposed components of our method on the R2R unseen dataset.

Analysis on the Dual Semantic-Aware Modules. Table 3 shows several ablation studies to verify the effectiveness of our proposed dual semantic-aware modules consisting of ASV and IGL. Specifically, ASV and IGL are proposed to improve the semantic representations hidden in vision and language, respectively. We can observe that ASV and IGL are beneficial to effectively enhance the performance on all metrics alone, and this improvement is much more pronounced when these two modules are combined, with gains of 3.3% on SR and 1.3% on SPL. This demonstrates that with the help of semantics augmentation on both modalities, the model is capable to understand environments and instructions better.

Analysis on the Augmented Memory Mechanism. To enhance memory representations through navigation steps, we propose GAA and RMF modules to adaptively aggregate the explicit global features of sub-images for each viewpoint, and transmit implicit reasoning states for each step, respectively. As shown in Table 4, for the graph memory (denoted as "GM"), GAA enhances all metrics significantly, demonstrating that our GAA method enables the model to assign weights reasonably by learning the different contributions of images based on their locations and contents. Additionally, for the recurrent states (denoted as "RS"), we explore the effects of



Move forward **past** the **exercise equipment** and **exit** the **room** through the doorway. Once outside the room, **turn left** and **enter** that **doorway**, then **wait** in that **room**.

(a)

Walk down the room on the **left** side. Stop in front of **the big windows** on the **left** looking out onto some **marble columns** outside.

(b)

Walk to the **piano**. Make **right**. **Wait** at the **door** next to the **piano**.

(c)

Figure 6: Visualization of navigation paths. The picture beside contains key guidance items which are marked in green. Below are the given language instructions, in which we use red for directions and blue for landmarks. Important phrases are highlighted by red boxes.

Method			SR↑	SPL↑	NE↓	OSR↑
GM	w/o GAA		71.39	61.24	3.22	79.95
	w/ GAA		**72.50**	**61.56**	**3.00**	**80.97**
RS	w/o RMF		72.16	60.73	3.14	80.29
	w/ RMF	w/o global	72.35	61.04	3.10	80.50
		w/o local	71.18	60.89	3.13	80.25
		w/o text	70.63	60.34	3.25	78.67
		Full	**72.50**	**61.56**	**3.00**	**80.97**

Table 4: Ablation study for the memory mechanism.

different branches on the RMF. It shows that the reasoning clues are simultaneously dependent on global-local features and language instructions. Overall, the agent can perform and leverage the history-dependent inference capabilities more effectively with the proposed GAA and RMF modules.

Analysis on the Number of Objects. As described in Sec. 3.2, the cross-image relationships of objects are first learned by a transformer encoder block. Therefore, the maximum number of objects in each image will influence the lengths of tokens. Fig. 7 shows that the model learning is optimal when the number of objects is 3 (totaling 108 objects for the current node). This demonstrates that too many objects will make the parallel lengths too long to pay attention to leading object features, while too few objects will result in a misinterpretation of visual semantics.

4.4 Qualitative Results and Visualization

Some visualization results are presented in this subsection to further analyze how our model contributes to the action decision during navigation. From Fig. 6 (a) we can see that our DSRG successfully seizes the moment to "turn left", that is, "once outside the room". However, DUET turns left late and enters the bathroom incorrectly. This indicates that our DSRG can better recognize the navigation progress and response for the explicit landmark. As for (b) and (c), our DSRG successfully chooses the right navigable nodes with respect to the "big windows on the left" and the "walk to the piano", which leads to correct navigation results, while DUET fails. This proves that a deeper understanding of the semantics both in instructions and observations is crucial for the VLN agent.

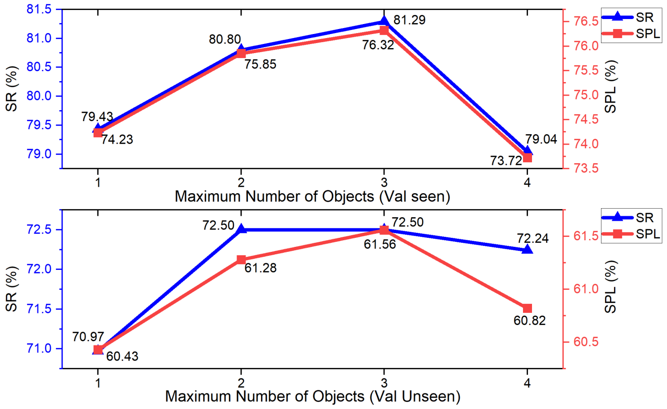


Figure 7: Ablation study for the maximum number of objects.

5 Conclusion and Future Work

In this paper, we present a dual semantic-aware recurrent global-adaptive network, namely DSRG, to improve the performance of agents in VLN. It can effectively recognize the guiding semantic information hidden in both linguistic instructions and visual observations with the help of the proposed ASV and IGL modules, respectively. For the memory mechanism, the GAA module is proposed to adaptively aggregate different sub-images in the panorama for global map construction. The RMF module is devised to supply implicit temporal hidden states by transferring reasoning cues from previous steps. Extensive experiments on two public datasets, R2R and REVERIE, have demonstrated that our method outperforms the state-of-the-art methods. With good expansibility and robustness, our approach is believed to have the potential to serve other VLN-like tasks as well, and we leave this exploration for future work.

Acknowledgments

This paper is supported by the National Natural Science Foundation of China under Grants (62233013, 62073245, 62173248). Suzhou Key Industry Technological Innovation-Core Technology R&D Program (SGC2021035); Special funds for Jiangsu Science and Technology Plan (BE2022119). Shanghai Municipal Science and Technology Major Project (2021SHZDZX0100) and the Fundamental Research Funds for the Central Universities. Shanghai Science and Technology Innovation Action Plan (22511104900).

References

[An *et al.*, 2021] Dong An, Yuankai Qi, Yan Huang, Qi Wu, Liang Wang, and Tieniu Tan. Neighbor-view enhanced model for vision and language navigation. In *ACMMM*, pages 5101–5109, 2021.

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

[Bird *et al.*, 2009] Steven Bird, Ewan Klein, and Edward Loper. *Natural language processing with Python: analyzing text with the natural language toolkit.* " O'Reilly Media, Inc.", 2009.

[Chang *et al.*, 2017] Angel Chang, Angela Dai, Thomas Funkhouser, Maciej Halber, Matthias Niessner, Manolis Savva, Shuran Song, Andy Zeng, and Yinda Zhang. Matterport3D: Learning from RGB-D data in indoor environments. *3DV*, 2017.

[Chen *et al.*, 2021a] Kevin Chen, Junshen K Chen, Jo Chuang, Marynel Vázquez, and Silvio Savarese. Topological planning with transformers for vision-and-language navigation. In *CVPR*, pages 11276–11286, 2021.

[Chen *et al.*, 2021b] Shizhe Chen, Pierre-Louis Guhur, Cordelia Schmid, and Ivan Laptev. History aware multimodal transformer for vision-and-language navigation. *NeurIPS*, 34, 2021.

[Chen *et al.*, 2022a] Jinyu Chen, Chen Gao, Erli Meng, Qiong Zhang, and Si Liu. Reinforced structured state-evolution for vision-language navigation. In *CVPR*, pages 15450–15459, 2022.

[Chen *et al.*, 2022b] Shizhe Chen, Pierre-Louis Guhur, Makarand Tapaswi, Cordelia Schmid, and Ivan Laptev. Learning from unlabeled 3d environments for vision-and-language navigation. In *ECCV*, pages 638–655. Springer, 2022.

[Chen *et al.*, 2022c] Shizhe Chen, Pierre-Louis Guhur, Makarand Tapaswi, Cordelia Schmid, and Ivan Laptev. Think global, act local: Dual-scale graph transformer for vision-and-language navigation. In *CVPR*, pages 16537–16547, 2022.

[Dang *et al.*, 2022a] Ronghao Dang, Zhuofan Shi, Liuyi Wang, Zongtao He, Chengju Liu, and Qijun Chen. Unbiased directed object attention graph for object navigation. In *ACMMM*, page 3617–3627, 2022.

[Dang *et al.*, 2022b] Ronghao Dang, Liuyi Wang, Zongtao He, Shuai Su, Chengju Liu, and Qijun Chen. Search for or navigate to? dual adaptive thinking for object navigation. *arXiv preprint arXiv:2208.00553*, 2022.

[Devlin *et al.*, 2019] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *NAACL*, pages 4171–4186. Association for Computational Linguistics, 2019.

[Dosovitskiy *et al.*, 2020] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[Fried *et al.*, 2018] Daniel Fried, Ronghang Hu, Volkan Cirik, Anna Rohrbach, Jacob Andreas, Louis-Philippe Morency, Taylor Berg-Kirkpatrick, Kate Saenko, Dan Klein, and Trevor Darrell. Speaker-follower models for vision-and-language navigation. *NeurIPS*, 31, 2018.

[Guhur *et al.*, 2021] Pierre-Louis Guhur, Makarand Tapaswi, Shizhe Chen, Ivan Laptev, and Cordelia Schmid. Airbert: In-domain pretraining for vision-and-language navigation. In *ICCV*, pages 1634–1643, 2021.

[Hao *et al.*, 2020] Weituo Hao, Chunyuan Li, Xiujun Li, Lawrence Carin, and Jianfeng Gao. Towards learning a generic agent for vision-and-language navigation via pretraining. In *CVPR*, pages 13137–13146, 2020.

[Hong *et al.*, 2021] Yicong Hong, Qi Wu, Yuankai Qi, Cristian Rodriguez-Opazo, and Stephen Gould. Vln bert: A recurrent vision-and-language bert for navigation. In *CVPR*, pages 1643–1653, 2021.

[Li *et al.*, 2022] Jialu Li, Hao Tan, and Mohit Bansal. Envedit: Environment editing for vision-and-language navigation. In *CVPR*, pages 15407–15417, 2022.

[Liang *et al.*, 2022a] Xiwen Liang, Fengda Zhu, Lingling Li, Hang Xu, and Xiaodan Liang. Visual-language navigation pretraining via prompt-based environmental self-exploration. In *ACL*, pages 4837–4851, 2022.

[Liang *et al.*, 2022b] Xiwen Liang, Fengda Zhu, Yi Zhu, Bingqian Lin, Bing Wang, and Xiaodan Liang. Contrastive instruction-trajectory learning for vision-language navigation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 1592–1600, 2022.

[Lin *et al.*, 2021] Xiangru Lin, Guanbin Li, and Yizhou Yu. Scene-intuitive agent for remote embodied visual grounding. In *CVPR*, pages 7036–7045, June 2021.

[Lin *et al.*, 2022] Bingqian Lin, Yi Zhu, Zicong Chen, Xiwen Liang, Jianzhuang Liu, and Xiaodan Liang. Adapt:

Vision-language navigation with modality-aligned action prompts. In *CVPR*, pages 15396–15406, 2022.

[Liu *et al.*, 2021] Chong Liu, Fengda Zhu, Xiaojun Chang, Xiaodan Liang, Zongyuan Ge, and Yi-Dong Shen. Vision-language navigation with random environmental mixup. In *ICCV*, pages 1644–1654, 2021.

[Ma *et al.*, 2019] Chih-Yao Ma, Zuxuan Wu, Ghassan Al-Regib, Caiming Xiong, and Zsolt Kira. The regretful agent: Heuristic-aided navigation through progress estimation. In *CVPR*, pages 6732–6740, 2019.

[Moudgil *et al.*, 2021] Abhinav Moudgil, Arjun Majumdar, Harsh Agrawal, Stefan Lee, and Dhruv Batra. Soat: A scene-and object-aware transformer for vision-and-language navigation. *NeurIPS*, 34:7357–7367, 2021.

[Qi *et al.*, 2020a] Yuankai Qi, Zizheng Pan, Shengping Zhang, Anton van den Hengel, and Qi Wu. Object-and-action aware model for visual language navigation. In *ECCV*, pages 303–317. Springer, 2020.

[Qi *et al.*, 2020b] Yuankai Qi, Qi Wu, Peter Anderson, Xin Wang, William Yang Wang, Chunhua Shen, and Anton van den Hengel. Reverie: Remote embodied visual referring expression in real indoor environments. In *CVPR*, pages 9982–9991, 2020.

[Qi *et al.*, 2021] Yuankai Qi, Zizheng Pan, Yicong Hong, Ming-Hsuan Yang, Anton van den Hengel, and Qi Wu. The road to know-where: An object-and-room informed sequential bert for indoor vision-language navigation. In *ICCV*, pages 1655–1664, 2021.

[Qiao *et al.*, 2022] Yanyuan Qiao, Yuankai Qi, Yicong Hong, Zheng Yu, Peng Wang, and Qi Wu. Hop: History-and-order aware pre-training for vision-and-language navigation. In *CVPR*, pages 15418–15427, 2022.

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

[Tan and Bansal, 2019] Hao Tan and Mohit Bansal. Lxmert: Learning cross-modality encoder representations from transformers. In *EMNLP-IJCNLP*, pages 5100–5111, 2019.

[Tan *et al.*, 2019] Hao Tan, Licheng Yu, and Mohit Bansal. Learning to navigate unseen environments: Back translation with environmental dropout. In *NAACL*, pages 2610–2621, 2019.

[Vaswani *et al.*, 2017] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NeurIPS*, pages 5998–6008, 2017.

[Wang *et al.*, 2019] Xin Wang, Qiuyuan Huang, Asli Celikyilmaz, Jianfeng Gao, Dinghan Shen, Yuan-Fang Wang, William Yang Wang, and Lei Zhang. Reinforced cross-modal matching and self-supervised imitation learning for vision-language navigation. In *CVPR*, pages 6629–6638, 2019.

[Wang *et al.*, 2020] Hanqing Wang, Wenguan Wang, Tianmin Shu, Wei Liang, and Jianbing Shen. Active visual information gathering for vision-language navigation. In *ECCV*, pages 307–322. Springer, 2020.

[Wang *et al.*, 2021] Hanqing Wang, Wenguan Wang, Wei Liang, Caiming Xiong, and Jianbing Shen. Structured scene memory for vision-language navigation. In *CVPR*, pages 8455–8464, 2021.

[Wang *et al.*, 2023] Liuyi Wang, Zongtao He, Ronghao Dang, Huiyi Chen, Chengju Liu, and Qijun Chen. Resss: Referring expression speaker via self-training with scorer for goal-oriented vision-language navigation. *IEEE Transactions on Circuits and Systems for Video Technology*, 2023.

[Zhang *et al.*, 2021] Yubo Zhang, Hao Tan, and Mohit Bansal. Diagnosing the environment bias in vision-and-language navigation. In *IJCAI*, 2021.

[Zhao *et al.*, 2022] Yusheng Zhao, Jinyu Chen, Chen Gao, Wenguan Wang, Lirong Yang, Haibing Ren, Huaxia Xia, and Si Liu. Target-driven structured transformer planner for vision-language navigation. In *ACMMM*, pages 4194–4203, 2022.

[Zhu and Yang, 2020] Linchao Zhu and Yi Yang. Actbert: Learning global-local video-text representations. In *CVPR*, pages 8746–8755, 2020.

[Zhu *et al.*, 2020] Fengda Zhu, Yi Zhu, Xiaojun Chang, and Xiaodan Liang. Vision-language navigation with self-supervised auxiliary reasoning tasks. In *CVPR*, pages 10012–10022, 2020.