2026 IEEE International Conference on Robotics and Automation (ICRA 2026)
June 1-5, 2026. Vienna, Austria

Awaken Memories with Words: Recursive Visual Imagination and Adaptive Linguistic Grounding for Vision Language Navigation

Bolei Chen, Jiaxu Kang, Yifei Wang, Ping Zhong†, and Jianxin Wang

***Abstract*— Vision Language Navigation (VLN) typically requires agents to navigate to specified objects or remote regions in unknown scenes by obeying linguistic commands. Such tasks require organizing historical visual observations for linguistic grounding, which is critical for long-sequence navigational decisions. However, current agents suffer from overly detailed scene representation and ambiguous vision-language alignment, which weaken their comprehension of navigation-friendly high-level scene priors and easily lead to behaviors that violate linguistic commands. To tackle these issues, we propose a navigation policy by recursively summarizing along-the-way visual perceptions, which are adaptively aligned with commands to enhance linguistic grounding. In particular, by structurally modeling historical trajectories as compact neural grids, several Recursive Visual Imagination (RVI) techniques are proposed to motivate agents to focus on the regularity of visual transitions and semantic scene layouts, instead of dealing with misleading geometric details. Then, an Adaptive Linguistic Grounding (ALG) technique is proposed to align the learned situational memories with different linguistic components purposefully. Such fine-grained semantic matching facilitates the accurate anticipation of navigation actions and progress. Our navigation policy outperforms the state-of-the-art methods on the challenging VLN-CE and ObjectNav tasks, showing the superiority of our RVI and ALG techniques for VLN.**

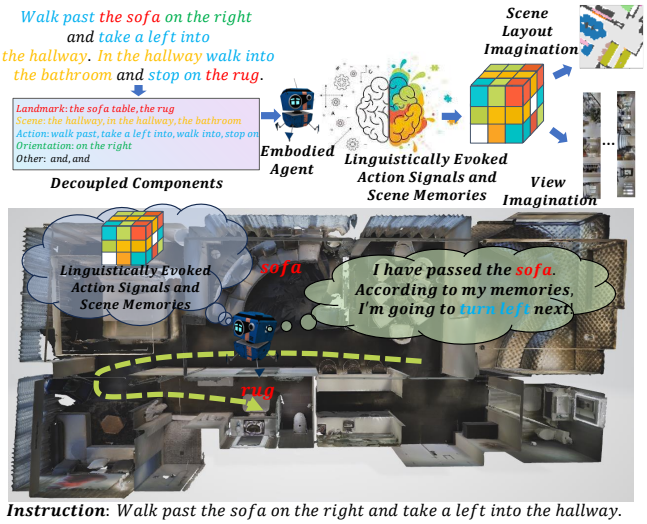

Fig. 1. The VLN agent decouples an instruction into different components, including landmarks, scenes, actions, orientations, and others, which are adaptively aligned with high-level scene priors in the ISR. The pre-trained ISR can provide the necessary mindsets for VLN, including view imagination and scene layout imagination.

I. Introduction

Interacting with agents through natural language is a long-term goal of embodied artificial intelligence as it is potentially the most intuitive way for human-robot communication. The emerging research on **V**ision **L**anguage **N**avigation (VLN) [1] is along this path, which requires agents to navigate to specified object instances or remote areas in unfamiliar 3D scenes by following linguistic instructions. Existing VLN work has made great advances in **S**cene **R**epresentation (SR) [1]–[3], vision-language alignment [4], and auxiliary tasks [5] for pre-training. They typically organize historical visual observations as structural SRs, which are further cross-modally aligned with instructions to track navigation progress and enhance navigation decision-making.

Some methods [2], [6] represent scenes by projecting raw or encoded visual features into bird's-eye-view maps or 3D feature fields to preserve fine-grained scene geometries and visual contexts. Despite promising progress has been made, these SRs provide overly detailed structural and semantic priors, posing challenges for learning accurate vision-action mappings using neural networks. Human-like agents typically establish high-level awareness of landmark semantics and spatial relationships of surrounding objects, rather than focusing on misleading geometric details that are irrelevant to navigation. For example, the agent in Fig. 1 should focus on the sofa landmarks and the visual signals that trigger the left-turn action, rather than the objects' visual textures and the hallway's geometric structure. Research in behavioral psychology [7] has shown that many animals maintain spatial representations of their scenes during navigation, even if scene details are not fully stored. Inspired by this, some other methods [1] propose to abstract the scene layouts into visual feature-based **T**opological **S**cene **R**epresentations (TSR) to facilitate linguistic grounding or balance exploration and exploitation during navigation. Although TSR refines the scene layout, TSR's nodes still store raw or encoded visual textures that are overly detailed. Moreover, TSR discards continuous semantic relations between nodes [8].

Redundant SRs can impede linguistic grounding, potentially resulting in behaviors contradicting navigation instructions. In other words, redundant scene details that are irrelevant to VLN can disrupt effective linguistic grounding, leading to ambiguous or even erroneous vision-language alignment. Current methods [1]–[3] attempt to align instruction tokens with SRs through standard cross-modal attention techniques. In this case, it is extremely challenging to train

This work was supported in part by the Key project of Xiangjiang Laboratory (25XJ02003) and in part by the National Natural Science Foundation of China under 62272489, 62332020, and 62350004. The authors are with the [1]School of Computer Science and Engineering, Central South University, Changsha 410083, China. Ping Zhong is also affiliated with [2]Xiangjiang Laboratory. † Corresponding Author (e-mail: ping.zhong@csu.edu.cn).

a transformer to achieve disentanglement and match each instruction token to the correct visual feature in a redundant SR. Such an ambiguous semantic alignment impairs the agent's insight into the navigation progress and makes it easy to deviate from the correct trajectory.

To tackle these issues, we propose a VLN policy by organizing along-the-way observations as an **I**mplicit **S**cene **R**epresentation (ISR) through **R**ecursive **V**isual **I**magination (RVI), including view imagination and scene layout imagination. Technically, we advocate modeling historical navigation trajectories (including the agent's visual sensing, poses, and navigational actions) as compact neural grids, rather than preserving explicit scene geometric details. We treat SR learning as a sequence modeling problem and train a joint state-action transformer over entire trajectories under the behavior cloning framework [9]. Unlike classical VLN methods [10], [11], the number of neural grids in our ISR is a hyperparameter that does not grow with trajectory length or scene scale. Therefore, the number of ISR tokens input to our model is fixed, which does not increase the computational cost. Then, the learned ISR is densely aligned with navigation commands via a novel **A**daptive **L**inguistic **G**rounding (ALG) technique to make the vision-language matching clear.

To derive navigation-friendly high-level scene priors from an ISR, RVI motivates agents to focus on the regularity of visual transitions and semantic scene layouts while ignoring irrelevant visual contexts. In particular, view imagination motivates agents to learn the distribution of future visual frames while enhancing their sensitivity to historical visual changes. Due to the inherent uncertainty in future frame prediction and the diversity of navigational actions, a single current frame can generate multiple potential futures. Therefore, our VLN agent is encouraged to summarize the regularity of visual signal changes instead of deterministically rendering future visual features. Scene layout imagination is designed to enhance the agent's insights into the surrounding landmark semantics and their relative positional relations. Therefore, our core idea is to explicitly endow the agent with the thinking necessary for VLN: **(1) recalling the past and predicting the future and (2) imagining the current semantic layout of the surroundings.**

Research in brain science [12] has shown that the cerebellum and hippocampus regulate motion and memory recall through neural structures and feature representations, respectively. Inspired by this, the ALG technique is proposed to adaptively align ISR's neural grids with different linguistic components for vision-language matching. For example, *left turn* action signals and *sofa* associated situational memories should be governed by separate neural grids, as shown in Fig. 1. To realize this idea, the agent first decouples an instruction into different components through syntactic analysis. Then, a self-supervised learning method is proposed to adaptively align these components with appropriate action signals or scene memories at the positional and semantic levels.

During experiments, sufficient comparative studies reflect that our approach incorporating RVI and ALG achieves state-of-the-art performance on two VLN tasks. Adequate ablation studies validate the effectiveness of the individual modules of our method. In general, the main contributions of this paper are as follows: **(1)** Two novel RVI techniques are designed for ISR learning that can empower agents with the essential thinking for VLN. **(2)** A novel ALG technique is proposed to motivate the agent to adaptively activate different action signals or scene memories based on different linguistic components. **(3)** Sufficient comparative and ablative studies on challenging VLN tasks demonstrate the superiority of our method.

II. Related Work

A. Scene Representation for VLN

Effective SRs are essential for the long-sequence decision-making and vision-instruction alignment of VLN. Early efforts [13], [14] typically employ recurrent neural networks to model SR as a fixed-size feature vector, which may be inefficient in modeling sophisticated visual features and capturing the long-term feature dependence in historical trajectories. Due to the strong expression power of transformer [9], transformer-based models [2], [4], [5], [15] have manifested their potential in VLN. Among them, architecture enhancement methods [15], [16] consider how to apply the powerful transformer structure to VLN under the reinforcement learning framework, facilitating more precise modeling of scenes. Trajectory optimization methods [2], [4], [5] treat VLN tasks as sequence modeling problems and train joint state-action models over entire trajectories under the behavior cloning framework.

Alternatively, some other methods [2], [11] achieve SR by projecting encoded visual features into egocentric semantic maps or topological graphs, which exhaustively retain the visual contexts and scene geometries. Although these methods achieve promising results, their SRs contain redundant information. We argue that SR should adequately represent the high-level scene-understanding mindsets required for VLN, rather than providing agents with excessive and misleading scene details. Inspired by the trajectory optimization methods [5], [17], we propose an ISR by modeling historical observations as compact neural grids. Unlike existing methods [2], [5], [16], we condense and refine the valuable historical information before feeding it into the cross-modal fusion module. In other words, the ISR is learned to emphasize the agent's insights into high-level visual signals and semantic scene layouts, which is distinct from existing SR modeling.

B. Linguistic Grounding for VLN

Fine-grained linguistic grounding is critical for instruction-following action prediction and VLN progress tracking. However, existing methods [1], [2], [18] coarsely align all instruction tokens with the SR at the sentence level, which impairs the agent's insight into the navigation progress. Some other studies [5] adopt auxiliary tasks to sequentially align historical observations with instructions during the pre-training phase. However, the positional and semantic alignments between historical observations and instruction tokens

are still ambiguous. To mitigate these issues, alternative methods [4] decouple navigation instructions into actions and landmarks and match them with entities in the panoramic images at a fine-grained level. However, given the diversity of scenes and the complexity of instructions, it is inadequate to bridge the vision-language gap using only navigational actions and entity landmarks.

To address the above issues, we propose to decouple a navigation instruction into different components, including landmarks, scenes, actions, and orientations. Then, an ALG technique is proposed to achieve dense alignment between the linguistic components and the ISR at the positional and semantic levels, respectively. The ALG technique allows VLN agents to evoke different episodic memories adaptively according to different linguistic components.

III. Preliminaries

A. Problem Definition

In this work, we address the VLN tasks in 3D indoor scenes, where the agents are required to reach specified remote regions or object instances. In particular, we focus on two practical settings: VLN in **C**ontinuous **E**nvironments (VLN-CE) [19] and **Object**-goal **Nav**igation (ObjectNav) [20] tasks in continuous scenes, where the agents should take low-level navigational actions. The action space consists of a set of parameterized discrete actions, e.g., *Forward (0.25m)*, *Turn Left/Right (15°)*, and *Stop*. Both VLN-CE and ObjectNav utilize the Habitat simulator [21] to render RGB and depth observations based on the MatterPort3D (MP3D) [22] dataset. In addition, the agents can receive noiseless 3-DoF pose data (x, y, θ), including 2D position and 1D orientation. At timestep t, the VLN agent can observe panoramic RGB images $\mathcal{R}_t = \{I_{t,k}^{rgb}\}_{k=1}^{K}$ and depth images $\mathcal{D}_t = \{I_{t,k}^{depth}\}_{k=1}^{K}$ of its current location, which both contain K single view images. The VLN agent also receives an instruction with L words for each episode, which are embedded as $X = \{x_i\}_{i=1}^{L}$. The ObjectNav agent can observe one single RGB image I_t^{rgb} and one single depth image I_t^{depth}. In each episode, the ObjectNav agent is given a target category c_{target} specified by a semantic label (e.g., a toilet). To facilitate the learning of a unified VLN framework, ObjectNav's goal is converted to "*Please navigate to [c_{target}] and stay within 1 m of it.*" by using a fixed instruction template. Unless otherwise stated, we default to introducing our method under the VLN setup.

B. ISR Initialization and Updating

At timestep t, the agent's observations specifically include the panoramic RGB-D images $\{\mathcal{R}_t, \mathcal{D}_t\}$, the pose (x_t, y_t, θ_t), and the previous navigation action a^{t-1}, as shown in Fig. 2. Following existing work [1], [2], [23], we first perform orientation embedding for each view of the panoramic image. Then, the pre-trained CLIP ResNet50 and the ResNet18 pre-trained in PointNav [24] are used to encode the individual RGB view $I_{t,k}^{rgb}$ and depth view $I_{t,k}^{depth}$, respectively. Notably, the visual encoders stay frozen

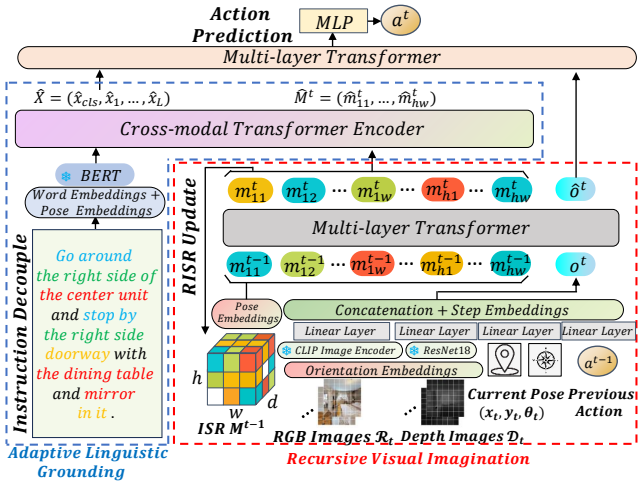

Fig. 2. An illustration of our VLN policy with RVI (Fig. 3 (a)) and ALG (Fig. 3 (b)). Our method treats SR learning as a sequence modeling problem and trains a joint state-action transformer over entire trajectories.

to make the training efficient. The agent's current pose is converted into a vector $(x_t, y_t, sin\theta_t, cos\theta_t)$ before encoding. Four different linear layers are used to project the visual embeddings, the pose vector, and the previous action into the same dimension. All the features are concatenated and further added a sinusoidal positional embedding of timestep t to obtain the current observation feature o^t.

Our ISR summarizes the historical images until timestep t as neural grids $M^t = [m_{ij}^t]_{h \times w}$ with $h \times w$ grids. Each grid is a d-dimensional feature vector $m_{ij}^t \in \mathbb{R}^d$ whose position with respect to the center is designated $[i - h/2, j - w/2]$. As each episode starts, the neural grids M^0 are initialized using their positions $m_{ij}^0 = w_m^0 + MLP([i - h/2, j - w/2])$, where $w_m^0 \in \mathbb{R}^d$ is a learnable embedding. At each timestep, the neural grids are updated given the new observation o^t with a differentiable function. Given the effectiveness of transformers in sequential modeling and VLN [10], a multi-layer transformer is employed to achieve interactions among neural grid-based situational memories. We first perform positional embedding for neural grids to enhance the geometry alignment between the neural grids and the observation. Then, all the neural grids and o^t are concatenated as tokens which are fed to the transformer, as shown in Fig. 2.

Notably, unlike the voxels for 3D scene reconstruction, we introduce the concept of a "grid" to emphasize the relative positional encoding of ISR. In the following section, we expect agents to predict local semantic maps during RVI, which requires inferring the relative positional relations between high-level semantics. In addition, we expect the grids with different positions to be aligned with the corresponding instruction components during ALG. This is inspired by the fact that the hippocampus and cerebellum, which have different relative positions in the brain, are responsible for memory and movement, respectively.

IV. Methodology

A. Recursive Visual Imagination

To derive high-level scene priors from ISR, RVI motivates agents to focus on the regularity of visual transitions

and semantic scene layouts while ignoring irrelevant visual contexts. As shown in Fig. 3 (a), RVI specifically includes **V**iew **I**magination (VI), **S**cene **L**ayout **I**magination (SLI), and **V**isual **S**emantic **P**rediction (VSP).

Given a query pose, VI motivates the agent to evoke the corresponding situational memory from ISR or learn the regularity of future visual transitions. At timestep t, we randomly sample a query pose $\{x_{t'}, y_{t'}, \theta_{t'}\}$ and the corresponding RGB panoramic image $\mathcal{R}_{t'}$ from a VLN trajectory, where $t' \in [0, t+k]$. Then, a frozen pre-trained CLIP ResNet50 and a linear layer are utilized to encode $\mathcal{R}_{t'}$ and the query pose as $v_{t'}$ and $q_{t'}$, respectively. As shown in Fig. 3 (a), $q_{t'}$ is fed into the multi-layer transformer along with M^{t-1} and o^t to query visual features about pose $\{x_{t'}, y_{t'}, \theta_{t'}\}$ from ISR. Notably, we only aim to extract potential features related to the query pose from the ISR, without expecting $q_{t'}$ to affect the ISR updating. Therefore, an attention masking operation is employed to prevent M^{t-1} and o^t from paying attention to $q_{t'}$. The output pose embedding is fed into an **M**ulti-**L**ayer **P**erception (MLP) to predict the visual feature $v^q_{t'}$. To enhance the agent's sensitivity to historical visual changes, we use a contrastive loss to clarify the correspondence between the poses and visual features by pushing $v^q_{t'}$ and $v_{t'}$ closer to each other and moving $v^q_{t'}$ away from visual features at other locations in the trajectory:

$$\mathcal{L}_{Con} = \frac{1}{T}\sum_{t=0}^{T} -log\frac{exp(sim(v^q_{t'}, v_{t'})/\tau)}{\sum_{i=1}^{t+k} exp(sim(v^q_{t'}, v_i)/\tau)}, \quad (1)$$

where τ is a softmax temperature scaling parameter and $sim(\cdot,\cdot)$ corresponds to the cosine similarity.

Notably, by setting the value of $k > 0$, the agent is motivated to imagine visual features for the future k timesteps at specific locations. To make the agent further summarize the regularity of future visual transitions, we aim to learn the distribution of future frames conditional on the current frame, rather than deterministically rendering future visual features. In particular, we employ two MLPs p_ϑ and q_ϑ to approximate the learned prior distribution $z_{t'} \sim p_\vartheta(z_{t'}|v^q_{t'})$ and the posterior distribution $\hat{z}_{t'} \sim q_\vartheta(\hat{z}_{t'}|v^q_{t'}, v_{t'})$ that captures future uncertainty, respectively. We make the prior distribution to be closer to the posterior distribution by minimizing the KL divergence, which not only enables the agent to fantasize about the future but also makes the future variable more predictable. In summary, the loss function for visual imagination is as follows:

$$\mathcal{L}_{VF} = \mathcal{L}_{Con} + \beta KL[q_\vartheta(z_{t'}|v^q_{t'}, v_{t'})||p_\vartheta(z_{t'}|v^q_{t'})], \quad (2)$$

where β is a loss scale hyperparameter. When $0 < t' \leq t$, $\beta = 0$, otherwise $\beta = 0.5$ $(t < t' \leq t+k)$.

SLI is designed to enhance the agent's insights into the surrounding landmark semantics and the relative positional relationships among them. Technically, an MLP is used to predict egocentric local semantic maps $\{\mathcal{M}^t\}_{t=0}^T$ from ISR, where $\{\mathcal{M}^t\}_{t=0}^T$ is pre-generated from the MP3D dataset, as shown in Fig. 3 (a). Please see Fig. 4 (b) for more details of $\mathcal{M}^t \in \mathbb{R}^{H\times W}$. A **B**inary **C**ross-**E**ntropy (BCE) loss is used to measure the SLI error:

$$\mathcal{L}_{Map} = \frac{1}{T}\sum_{t=0}^{T} BCE(Linear(M^t), \mathcal{M}^t). \quad (3)$$

To boost VI and SLI's focus on scene semantics, VSP is used as an auxiliary task to enhance the sensitivity of the observation encoding component to visual semantics. Technically, VSP is achieved to predict the existence of each object category and the ratio occupied by the objects in the current view (if they are present) based on the observation o^t, as shown in Fig. 3 (a). We obtain the ground-truth labels from the MP3D training scenes and use the BCE loss $\mathcal{L}_{Sem}$ to measure the VSP errors. More details regarding data collection for pretraining will be presented in the experimental section.

B. Adaptive Linguistic Grounding

Instruction Decoupling. Human beings can wisely focus on instruction-related landmarks in the scene and scene-related orientations in the instructions when performing VLN tasks. To emulate such abilities, we propose to decouple the instruction into different components, which are independently and adaptively aligned with ISR's neural grids, producing more discriminative and clear vision-language matching. Technically, we follow the existing work [25] to parse the instructions grammatically and decouple the instructions into five semantic components: landmarks, scenes, actions, orientations, and others. Particularly, we generate the position labels L_{land}, L_{scene}, L_{action}, L_{ori}, and L_{other} for the component's associated words by setting each component's word positions to 1 and the rest to 0, as shown in Fig. 3 (b). Given that large language models [26] can potentially solve this issue better, we report the related experimental results in diagnostic studies. In addition, by dot-multiplying the cross-modal fused word tokens $\{\hat{x}_i\}_{i=0}^L$ with the position labels, we derive the textual features of the decoupled components $\{\tilde{x}_i\}_{0<i\leq L}$. Notably, the decoupled textual features, as a result of cross-attention, implicitly contain information about the global instruction and ISR while preserving the original textual features. That is, feature decoupling produces individual features while keeping the global context.

VLN Progress Tracking. Since VLN's decision-making is progressive, the agent needs to track the navigation progress and explicitly align the already executed instruction components, rather than the entire instruction, with the ISR. As shown in Fig. 3 (b), an MLP is used to map the cross-modal fused tokens $\hat{X} = \{\hat{x}_1, ..., \hat{x}_L\}$ to instruction weights $W_t = [\omega^t_1, ..., \omega^t_L]$, which assign higher attention to the already executed instruction components. The training target d_t of progress tracking is defined as the normalized distance from the current viewpoint to the goal, i.e., the target will be 1 at the beginning and closer to 0 as the agent approaches the goal. We employ a mean squared loss $\mathcal{L}_{Pro}$ to supervise the training of the progress tracking module.

Position and Semantic Alignments. Before performing the ALG, we need to specify which neural grids are aligned with which components in the instruction. To this end, we

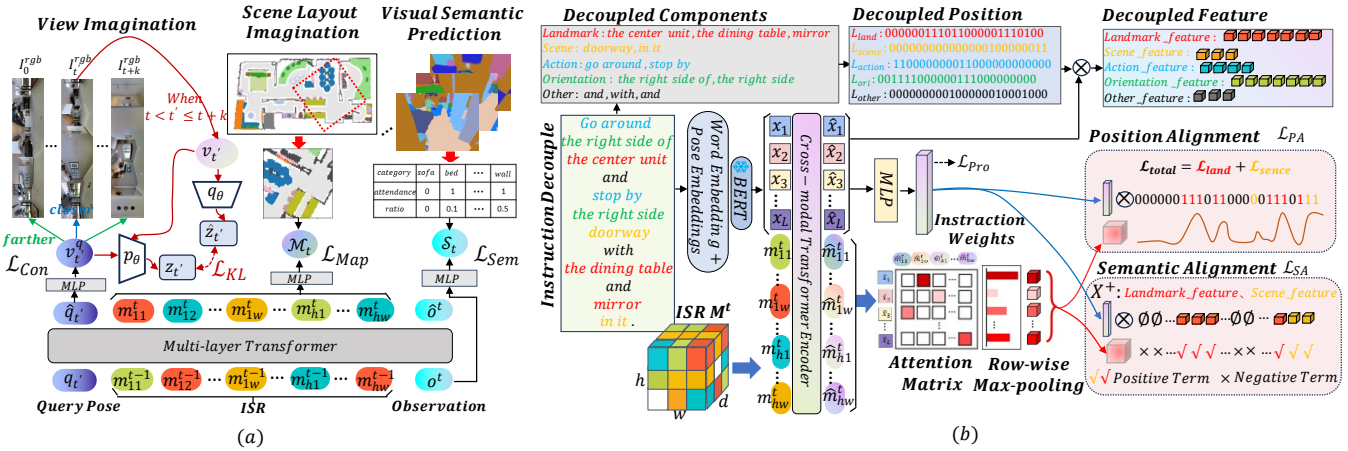

Fig. 3. (a) An illustration of RVI, including view imagination, scene layout imagination, and visual semantic prediction. (b) An illustration of ALG, including instruction decoupling, VLN progress tracking, and linguistic alignment.

propose to treat the attention matrix of the last cross-modal attention layer as an affinity matrix to match the neural grids and instruction components (as shown in Fig. 3 (b), since it is learned to adaptively measure the semantic similarity between the tokens [27]. Such an idea has two benefits: **(1)** No additional matching algorithms are required. **(2)** Such a design facilitates the agent to learn neural grid's adaptive attention to different instruction components when the model parameters are updated. Specifically, we first perform row-wise max-pooling on the attention matrix to obtain each language token's most attentive neural grid $\{\tilde{m}_i^t\}_{0<i\leq L}$. Note that $i \leq L$ since multiple language tokens pay attention to the same neural grid. Fig. 3 (b) shows an example of ISR actively and adaptively focusing on landmarks, scenes, i.e., positionally and semantically aligning $\{\tilde{m}_i^t\}_{0<i\leq L}$ with the landmark and scene components in the instruction. Those tokens that do not actively pay attention to landmarks and scenes are forced to match other instruction components, i.e., actions, orientations, and others. For brevity, only the ALG technique for landmark and scene alignment shown in Fig. 3 (b) is detailed below.

Position alignment aims to closely match the distribution of linguistically modulated ISR with the text distribution of instructions. The ground-truth text distribution of landmarks and scenes is obtained by element-wise summing the position labels of the associated decoupled text components, i.e., $L_{total} = L_{land} + L_{scene}$. In practice, we dot-multiply L_{total} and W_t to produce a ground-truth text distribution with navigational progress awareness, as shown in Fig. 3 (b). The process of position label prediction is as follows:

$$\hat{L}_{total} = Softmax(MLP(Mean([\tilde{m}_0^t, ..., \tilde{m}_i^t]))), \quad (4)$$

where $Mean(\cdot)$ denotes averaging over the neural grids. We use a BCE loss $\mathcal{L}_{PA}$ to supervise the training of the position alignment. Semantic alignment aims to match semantically similar neural grids with instruction components and keep away the dissimilar ones from both through contrastive learning. The semantic alignment loss is defined as follows:

$$\mathcal{L}_{SA} = \frac{1}{|X^+|} \sum_{\tilde{x}_i \in X^+} -log \frac{exp(\alpha_+ * (\overline{m}^\top \tilde{x}_i / \tau))}{\sum_{j=1}^{l} exp(\alpha_- * (\overline{m}^\top \tilde{x}_j / \tau))}, \quad (5)$$

where $X^+ = \{\tilde{x}_i\}_{0<i\leq L}$ denotes the text features corresponding to the landmark and scene components, as shown in Fig. 3 (b). l denotes the number of tokens in X^+ and $\overline{m} = Mean([\tilde{m}_0^t, ..., \tilde{m}_i^t])$. τ is a temperature scaling parameter. α_+ and α_- are the weights of positive term (landmarks and scenes) and negative term (actions, orientations, and others), respectively. Conversely, we can also utilize the ALG technique in practice to make agents actively and adaptively focus on action and orientation components in the instruction. Those tokens that do not actively pay attention to actions and orientations are forced to align with other instruction components, i.e., landmarks, scenes, and others. We will report the navigation performance of different variants in the ablation study section.

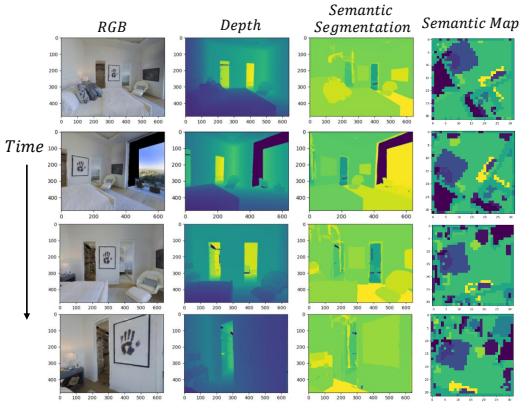

Fig. 4. Examples of observation sequences collected along the trajectory in a navigation episode. Only one view per timestep is shown here.

C. Pre-training and Fine-tuning for VLN

In the pre-training phase, we train the agent using a large number of pre-collected trajectories in the behavioral cloning framework [9]. A cross-entropy loss with inflection weighting [28] is employed for action prediction, which gives higher weights for actions different from the previous one:

$$\mathcal{L}_{Action} = \frac{1}{T} \sum_{t=0}^{T} -(1 + \gamma \mathbf{1}_{a_t^* \neq a_{t-1}^*} log(p(a_t^*))). \quad (6)$$

The total loss $\mathcal{L}_{total}$ in the pre-training phase is denoted as:

$$\mathcal{L}_{total} = \mathcal{L}_{Action} + \beta(\mathcal{L}_{VF} + \mathcal{L}_{Map} + \mathcal{L}_{Sem}) + \lambda(\mathcal{L}_{Pro} + \mathcal{L}_{PA} + \mathcal{L}_{SA}), \quad (7)$$

TABLE I

RESULTS ON THE R2R-CE DATASET.

Method	Val Unseen			Test Unseen		
	OSR↑	SR↑	SPL↑	OSR↑	SR↑	SPL↑
CM^2 [18]	42	34	28	39	31	24
GELA [4]	59	48	41	57	46	40
GridMM [2]	61	49	41	56	46	39
Ego^2-Map [3]	-	52	46	56	47	41
DREAMWALKER [23]	59	49	44	57	49	44
ETPNav [1]	65	57	49	63	55	48
Zhang et.al. [31]	-	58	49	-	56	48
Ours	**67**	**59**	**50**	**64**	**57**	**50**

TABLE II

RESULTS ON THE MP3D-OBJECTNAV DATASET (VAL).

Method	ObjectNav-MP3D (val)		
	SR(%)↑	SPL(%)↑	DTS(m)↓
Ego^2-MAP [32]	29.0	10.6	5.17
VLFM [6]	36.2	15.9	-
ECL [33]	34.8	14.7	4.95
T-Diff [34]	39.6	15.2	5.16
SG-Nav [35]	40.2	16.0	-
HOZ$_e$++ [20]	37.0	15.2	**4.11**
NaviFormer [36]	40.1	15.1	5.19
Ours	**40.9**	**17.1**	4.68

where β and λ are weighting parameters. Furthermore, the Dagger technique [29] is used to fine-tune the pre-trained models to address the distribution discrepancy between the offline training data and the target policy. Fine-tuning fundamentally differs from the pre-training phase that employs expert demonstration paths, as it involves novel data acquisition via exploration.

V. EXPERIMENTS

A. Experimental Settings and Implementation Details

Datasets and Evaluation Metrics. As stated in the problem definition, we evaluate our proposed VLN policy on the R2R-CE [19] and Habitat ObjectNav [30] datasets. For VLN, an episode is successful if the stop decision is taken within 3 m of the goal position. For ObjectNav, all goals are converted to instructions such as "*Please navigate to [c_{target}] and stay within 1 m of it.*" using a fixed instruction template. An episode is successful if the stop decision is made within 1 m of the object goal. There are several standard metrics [1] for VLN evaluation, including **S**uccess **R**ate (SR), **O**racle **SR** (OSR), and **S**R penalized by **P**ath **L**ength (SPL).

Implementation Details. The number of layers and attention heads of the transformers in our VLN strategy are 4 and 8, respectively. If not additionally specified, the dimensions of ISR are sized $h = w = 10$ and $d = 512$. τ and k in VI are respectively set to 0.07 and 20. All egocentric semantic maps used in SLI have a scale of $H = W = 32$ with each pixel corresponding to 20 $cm \times 20$ cm. The L in ALG is empirically set to 160 according to the length of the instructions in the R2R-CE dataset. The weights α_+, α_-, and τ in the semantic alignment of ALG are set to 1.0, 2.0, and 0.07, respectively. β and λ in Eq. 7 are set to 0.3 and 0.5. Following existing methods [1], [23], we employ a waypoint predictor for the VLN task to predict long-term navigation goals. For the ObjectNav task, we directly predict low-level navigation actions end-to-end.

For pre-training, we collect navigation trajectories based on the episodes in the training split, including visual observations, egocentric semantic maps, and semantically segmented views, as shown in Fig. 4. The whole model is trained for 100 epochs on one NVIDIA GeForce RTX 3090 GPU using a learning rate of 1×10^{-4} and batch size of 4. The optimizer is AdamW. For fine-tuning, our VLN policy is trained for more than 50 epochs on 4 NVIDIA GeForce RTX 3090 GPUs using a learning rate of 5×10^{-5} and 6 threads.

B. Comparison with State-of-the-art Methods

We first conduct comparative studies between our VLN policy and the state-of-the-art methods on the R2R-CE dataset. For adequate comparisons, the baselines are diverse in terms of SR. For example, CM^2, GridMM, and ETPNav employ the explicit semantic grid map, visual feature field, and TSR as SRs, respectively. Ego^2-Map uses a self-supervised SR learning scheme based on 2D-3D contrastive learning. However, these methods share the same drawback of using only cross-attention to ambiguously align SR with instruction features at the sentence level. GELA mitigates this problem and is similar to our ALG, but it only uses contrastive learning to align visual features with the object entities in the instructions. As shown in Tab. I, our method achieves the best performance on both splits, reflecting the superiority of our ISR and ALG techniques. Notably, DREAMWALKER attempts to learn a world model for predicting future views to augment VLN, which is different from our visual imagination. However, DREAMWALKER requires constructing an additional TSR, which is difficult to scale to large-scale scenes. Our method overcomes this issue by using ISR to organize historical images and imagine spatio-temporal high-level semantics, thus significantly outperforms DREAMWALKER.

As expected, our method also achieves the best performance on the ObjectNav dataset, as shown in Tab. II. Similarly, our method outperforms those methods that utilize semantic grid maps (HOZ$_e$++ and NaviFormer), visual feature fields (VLFM), and visual representations based on self-supervised contrastive learning (Ego^2-Map, and ECL). It is worth noting that T-Diff uses a trajectory diffusion technique to predict future trajectories, which is different from our idea of visual imagination. SG-Nav extracts commonsense knowledge from large language models to enhance ObjectNav, but relies on a TSR that are difficult to scale with scene size. Unlike the VLN methods in Tab. I, which predict navigational subgoals across multiple time steps, ObjectNav requires the agent to make navigational decisions at each time step, and thus relies more heavily on fine-grained vision-language alignment. To this end, our method has excellent visual imagination and ALG abilities, which significantly improve the ObjectNav performance.

C. Ablation Studies

We conduct ablation studies on the individual components of our method to clarify their contributions. All ablations utilize $\mathcal{L}_{Action}$ and $\mathcal{L}_{Sem}$ to ensure basic action prediction and effective observation encoding. As shown in Tab. III, all the RVI techniques ($\mathcal{L}_{Map}$, $\mathcal{L}_{Con}$, and $\mathcal{L}_{KL}$) can improve the VLN performance. In addition, the involvements of posi-

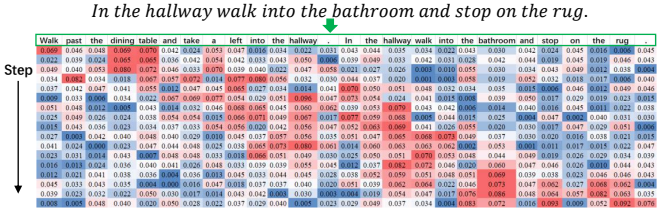

Fig. 5. illustrates how the instruction weights change with navigation progress. Different rows indicate weights at different time steps. A redder color indicates that the agent is more attentive to the corresponding words.

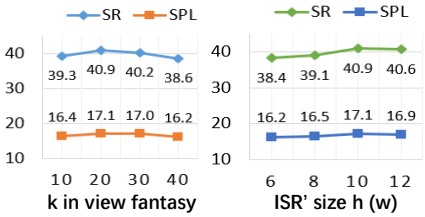

Fig. 6. Illustrations of parametric studies.

TABLE III

ABLATION STUDIES ON THE R2R-CE DATASET.

Ablations						Val Unseen		
$\mathcal{L}_{Map}$	$\mathcal{L}_{Con}$	$\mathcal{L}_{KL}$	$\mathcal{L}_{Pro}$	$\mathcal{L}_{PA}$	$\mathcal{L}_{SA}$	OSR ↑	SR ↑	SPL ↑
✗	✗	✗	✗	✗	✗	58	49	43
✓	✗	✗	✗	✗	✗	60	51	45
✓	✓	✗	✗	✗	✗	62	52	45
✓	✓	✓	✗	✗	✗	63	53	47
✓	✓	✓	✓	✓	✗	64	55	48
✓	✓	✓	✗	✓	✓	63	54	46
✓	✓	✓	✓	✓	✓	**67**	**58**	**50**

TABLE IV

VLN PERFORMANCE USING DIFFERENT ALG VARIANTS.

DIA Variants	Val Unseen		
	OSR ↑	SR ↑	SPL ↑
Action Priority	65	55	47
Scene Priority	67	58	50

tional alignment $\mathcal{L}_{PA}$ and semantic alignment $\mathcal{L}_{SA}$ promote ALG, which further leads to substantial OSR, SR, and SPL boosts. Notably, $\mathcal{L}_{PA}$ and $\mathcal{L}_{SA}$ should be used in conjunction with $\mathcal{L}_{Pro}$ as the navigation process is progressive. The absence of progress tracking $\mathcal{L}_{Pro}$ will result in a significant decrease in performance.

As shown in Fig. 3 (b), our proposed adaptive position and semantic alignments force ISR to actively focus on the landmark and scene components in the instructions, which we call *scene priority*. Alternatively, we can also design an action-aware ALG variant to motivate ISR to actively pay attention to the action and orientation components, which we call *action priority*. The comparative results in Tab. IV quantitatively evaluate the performance of two ALG variants. We find that the focus on scene and landmark components produces more efficient VLN agents under the R2R-CE setting. In other words, agents in the R2R-CE setup are more sensitive to landmark entities and scene references.

D. Diagnostic Studies and Discussion

(1) Does the VLN progress tracking work? Fig. 5 illustrates how the instruction weights change in the process tracking module as the VLN progresses. We find that the instruction weights in the progress tracking module can reflect which part of the instruction has been executed. In addition,

TABLE V

EFFECTS OF DIFFERENT INSTRUCTION DECOUPLING METHODS ON THE VLN PERFORMANCE ON THE R2R-CE DATASET.

Method	Val Unseen			Test Unseen		
	OSR ↑	SR ↑	SPL ↑	OSR ↑	SR ↑	SPL ↑
w/o manual check	66	58	49	63	57	50
w/ manual check	67	59	50	64	57	50
w/ GPT-4	67	60	51	65	58	50

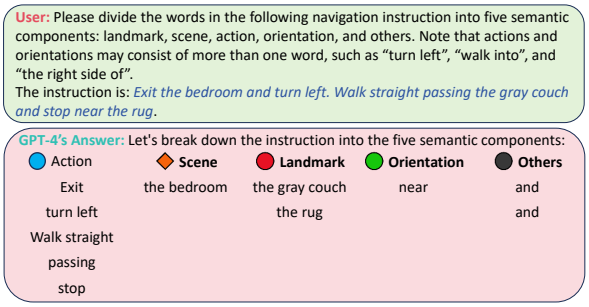

Fig. 7. An illustration of semantic component division based on GPT-4.

the instruction weights also reflect the agent's attention to the scene and landmark components of the instruction.

(2) How much does the hyperparameters affect our method? Fig. 6 illustrates the sensitivity analysis results for two key hyperparameters, i.e., the range of visual imagination (k), and the dimensions of ISR (h and w). For k, we evaluated four cases with $k = \{10, 20, 30, 40\}$. For h and w, we evaluated the four cases $h = w = \{6, 8, 10, 12\}$. We find that our method performs best when $k = 20$ and $w = h = 10$. In addition, our method is insensitive to these hyperparameters and thus is robust.

(3) Can instruction decoupling based on large language models achieve better performance? We use different instruction parsing schemes to decouple the navigation instructions in the R2R-CE dataset and investigate their effects on the VLN performance, the results are shown in Tab. V. The first row in Tab. V indicates that only off-the-shelf tools are used for instruction parsing without manual checks. The second row indicates the addition of a manual check. The third row indicates directly using the components decoupled by GPT-4, as shown in Fig. 7. The results show that GPT-4-based instruction decoupling leads to better VLN performance due to the powerful language analysis capability of large language models. When manual checking is missing, the decrease in VLN performance reflects the necessity of accurate instruction decoupling for positional and semantic alignments in the ALG.

VI. CONCLUSION

This paper focuses on scene representation and instruction grounding problems in VLN tasks. For scene representation, we enable the agent's abilities to model the regularity of visual transitions and semantic scene layouts by learning an ISR, rather than retaining redundant geometric details. In other words, we advocate empowering VLN agents with two necessary abilities: **(1) recalling the past and predicting the future and (2) imagining the current semantic layout of the surroundings.** For linguistic grounding, we suggest adaptively aligning the ISR with different instruction

components at the positional and semantic levels, rather than ambiguous vision-language matching. Sufficient comparative and ablation studies demonstrated our method's feasibility and superiority over existing methods. In the future, we will try to make efforts on zero-shot VLN based on multimodal large models to improve the generalization of VLN agents.

REFERENCES

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

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

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

[4] Y. Cui, L. Xie, Y. Zhang, M. Zhang, Y. Yan, and E. Yin, "Grounded entity-landmark adaptive pre-training for vision-and-language navigation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 12043–12053, 2023.

[5] S. Wu, X. Fu, F. Wu, and Z.-J. Zha, "Vision-and-language navigation via latent semantic alignment learning," *IEEE Transactions on Multimedia*, 2024.

[6] N. Yokoyama, S. Ha, D. Batra, J. Wang, and B. Bucher, "Vlfm: Vision-language frontier maps for zero-shot semantic navigation," in *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 42–48, IEEE, 2024.

[7] E. C. Tolman, "Cognitive maps in rats and men.," *Psychological review*, vol. 55, no. 4, p. 189, 1948.

[8] B. Chen, J. Kang, P. Zhong, Y. Cui, S. Lu, Y. Liang, and J. Wang, "Think holistically, act down-to-earth: A semantic navigation strategy with continuous environmental representation and multi-step forward planning," *IEEE Transactions on Circuits and Systems for Video Technology*, 2023.

[9] S. Hu, L. Shen, Y. Zhang, Y. Chen, and D. Tao, "On transforming reinforcement learning with transformers: The development trajectory," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.

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

[11] L. Wang, Z. He, J. Tang, R. Dang, N. Wang, C. Liu, and Q. Chen, "A dual semantic-aware recurrent global-adaptive network for vision-and-language navigation," *arXiv preprint arXiv:2305.03602*, 2023.

[12] A. A. Sokolov, R. C. Miall, and R. B. Ivry, "The cerebellum: adaptive prediction for movement and cognition," *Trends in cognitive sciences*, vol. 21, no. 5, pp. 313–332, 2017.

[13] R. Dang, Z. Shi, L. Wang, Z. He, C. Liu, and Q. Chen, "Unbiased directed object attention graph for object navigation," in *Proceedings of the 30th ACM International Conference on Multimedia*, pp. 3617–3627, 2022.

[14] S. Tan, K. Sima, D. Wang, M. Ge, D. Guo, and H. Liu, "Self-supervised 3-d semantic representation learning for vision-and-language navigation," *IEEE Transactions on Neural Networks and Learning Systems*, 2024.

[15] C. Lin, Y. Jiang, J. Cai, L. Qu, G. Haffari, and Z. Yuan, "Multimodal transformer with variable-length memory for vision-and-language navigation," in *European Conference on Computer Vision*, pp. 380–397, Springer, 2022.

[16] Y. Hong, Q. Wu, Y. Qi, C. Rodriguez-Opazo, and S. Gould, "Vln bert: A recurrent vision-and-language bert for navigation," in *Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition*, pp. 1643–1653, 2021.

[17] K. Ehsani, T. Gupta, R. Hendrix, J. Salvador, L. Weihs, K.-H. Zeng, K. P. Singh, Y. Kim, W. Han, A. Herrasti, *et al.*, "Spoc: Imitating shortest paths in simulation enables effective navigation and manipulation in the real world," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16238–16250, 2024.

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

[19] J. Krantz, E. Wijmans, A. Majumdar, D. Batra, and S. Lee, "Beyond the nav-graph: Vision-and-language navigation in continuous environments," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXVIII 16*, pp. 104–120, Springer, 2020.

[20] S. Zhang, X. Song, X. Yu, Y. Bai, X. Guo, W. Li, and S. Jiang, "Hoz++: Versatile hierarchical object-to-zone graph for object navigation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025.

[21] S. K. Ramakrishnan, A. Gokaslan, E. Wijmans, O. Maksymets, A. Clegg, J. Turner, E. Undersander, W. Galuba, A. Westbury, A. X. Chang, *et al.*, "Habitat-matterport 3d dataset (hm3d): 1000 large-scale 3d environments for embodied ai," *arXiv preprint arXiv:2109.08238*, 2021.

[22] A. Chang, A. Dai, T. Funkhouser, M. Halber, M. Niessner, M. Savva, S. Song, A. Zeng, and Y. Zhang, "Matterport3d: Learning from rgb-d data in indoor environments," *arXiv preprint arXiv:1709.06158*, 2017.

[23] H. Wang, W. Liang, L. Van Gool, and W. Wang, "Dreamwalker: Mental planning for continuous vision-language navigation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 10873–10883, 2023.

[24] E. Wijmans, A. Kadian, A. Morcos, S. Lee, I. Essa, D. Parikh, M. Savva, and D. Batra, "Dd-ppo: Learning near-perfect pointgoal navigators from 2.5 billion frames," *arXiv preprint arXiv:1911.00357*, 2019.

[25] H. Wu, J. Mao, Y. Zhang, Y. Jiang, L. Li, W. Sun, and W.-Y. Ma, "Unified visual-semantic embeddings: Bridging vision and language with structured meaning representations," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6609–6618, 2019.

[26] J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. L. Aleman, D. Almeida, J. Altenschmidt, S. Altman, S. Anadkat, *et al.*, "Gpt-4 technical report," *arXiv preprint arXiv:2303.08774*, 2023.

[27] A. Pardyl, G. Rypeść, G. Kurzejamski, B. Zieliński, and T. Trzciński, "Active visual exploration based on attention-map entropy," in *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence*, IJCAI '23, 2023.

[28] E. Wijmans, S. Datta, O. Maksymets, A. Das, G. Gkioxari, S. Lee, I. Essa, D. Parikh, and D. Batra, "Embodied question answering in photorealistic environments with point cloud perception," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6659–6668, 2019.

[29] S. Ross, G. Gordon, and D. Bagnell, "A reduction of imitation learning and structured prediction to no-regret online learning," in *Proceedings of the fourteenth international conference on artificial intelligence and statistics*, pp. 627–635, JMLR Workshop and Conference Proceedings, 2011.

[30] T. Gervet, S. Chintala, D. Batra, J. Malik, and D. S. Chaplot, "Navigating to objects in the real world," *Science Robotics*, vol. 8, 2022.

[31] Y. Zhang and P. Kordjamshidi, "Narrowing the gap between vision and action in navigation," in *Proceedings of the 32nd ACM International Conference on Multimedia*, pp. 856–865, 2024.

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

[33] B. Chen, J. Kang, P. Zhong, Y. Liang, Y. Sheng, and J. Wang, "Embodied contrastive learning with geometric consistency and behavioral awareness for object navigation," in *Proceedings of the 32nd ACM International Conference on Multimedia*, pp. 4776–4785, 2024.

[34] X. Yu, S. Zhang, X. Song, X. Qin, and S. Jiang, "Trajectory diffusion for objectgoal navigation," in *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

[35] H. Yin, X. Xu, Z. Wu, J. Zhou, and J. Lu, "Sg-nav: Online 3d scene graph prompting for llm-based zero-shot object navigation," *arXiv preprint arXiv:2410.08189*, 2024.

[36] W. Xie, H. Jiang, Y. Zhu, J. Qian, and J. Xie, "Naviformer: A spatio-temporal context-aware transformer for object navigation," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 39, pp. 14708–14716, 2025.