AHead-VLN: A Hierarchical Reasoning Framework Guided by Foresight for Human-Aware Navigation

Junyi Zhang[1,2], **Ziwen Luo**[1], **Yinzhe Zhou**[1], **Xinghe Chu**[1,2] and **Zhaoming Lu**[1*]
[1]Beijing University of Posts and Telecommunications
[2]Xiong'an Aerospace Information Research Institute
{zhangjy1112, luozw, zhouyinzhe, chuxinghe, lzy0372}@bupt.edu.cn

Abstract

In human-aware vision-and-language navigation (HA-VLN), agents need to follow instructions while navigating safely among people in crowded environments. Existing methods often struggle to effectively transform forward-looking risk prediction into decision-making basis, resulting in intelligent agents being prone to collisions in crowded scenarios and taking overly cautious actions, thereby reducing navigation efficiency. Moreover, it is difficult for high-level decisions based on social norms to be stably implemented in dynamic crowd scenarios, which can easily lead to behavioral jitter and planning deviation. In this work, we propose the AHead-VLN, a navigation framework that incorporates forward-looking risk information in the reasoning process. Specifically, we designed the path risk association module (PRA), which compiles future risks into structured risk events aligned with each candidate path, making risk information decision-ready for the large language model (LLM) reasoning. To balance executability and decision stability, we have further designed the hierarchical LLM planner (HLP) with two specialized components. The "Brain" LLM makes route choices based on instructions and risks, while the "Cerebellum" LLM prepares executable action scripts for candidate paths and retrieves the appropriate script for execution. Experiments on the HA-VLN benchmark show that AHead-VLN improves success while enhancing safety compared to strong baselines.

1 Introduction

Vision-and-Language Navigation (VLN) requires the agent to reach a target in an unknown environment using only first-person visual observations and natural language instructions [Anderson *et al.*, 2018; Gu *et al.*, 2022; Park and Kim, 2023]. Early VLN studies mainly focused on static environments where agents move between discrete viewpoints. However, real-world applications require intelligent agents to navigate in a human-centered dynamic environment and abide by social norms. This demand has driven the transition from traditional VLN to human-aware VLN, shifting the research focus from finding the shortest path to robust risk perception and decision-making in dynamic social environments [Möller *et al.*, 2021; Singamaneni *et al.*, 2024].

Figure 1: Navigation behaviours in a narrow corridor. (top) Reactive baseline: the agent relies only on a local cost map, blindly enters the narrow passage and gets stuck in a deadlock with an oncoming pedestrian. (down) AHead-VLN: the agent anticipates future crowd risk, chooses to wait at the entrance, and then proceeds, thus avoiding the deadlock.

Navigating in a dynamic social environment presents three major challenges. Firstly, the agent needs to predict and avoid collisions with pedestrians and other moving objects. Secondly, many critical risks lie outside the current field of view or are partially occluded, creating perceptual occlusion issues. Thirdly, the agent must respect implicit human norms of yielding, interpersonal distance, and passing order, which introduces social compliance requirements [Biswas *et al.*, 2022]. Thus, the problem is no longer "where to go" but "how and when to reach the goal", which exceeds the capacity of traditional geometric path planning [Singamaneni *et al.*, 2024; Zhang *et al.*, 2024a].

*Corresponding author.

The existing obstacle avoidance methods can be mainly classified into two types, but both have limitations. As shown in Figure 1(top), the reactive strategy was more common in the early VLN research. It relies on immediate control strategies and only makes short-sighted responses to the current state. In the face of complex situations, it is prone to deadlock and severe jitter [Paez-Granados *et al.*, 2022; Sathyamoorthy *et al.*, 2020; Gao and Huang, 2022]. The numerical prediction strategy incorporates time-related occupancy prediction and generates cost fields that change over time on the map [Thomas *et al.*, 2023; Poddar *et al.*, 2023; Le *et al.*, 2024]. Although it provides forward-looking predictions, it is only utilized by the underlying planning module and is difficult to support the intelligent agent in making a reasonable trade-off between efficiency and social norms, resulting in unnecessary detours and overly cautious actions [Xue *et al.*, 2024]. Therefore, risk information is introduced into the reasoning and decision-making process to serve as the basis for strategy weighing.

Large language models (LLMs) applied in VLN can provide powerful semantic understanding and common sense reasoning capabilities [Song *et al.*, 2023; Yao *et al.*, 2022]. However, directly embedding LLMs into embodied navigation brings additional challenges. Firstly, at the input level, LLMs can't directly understand the urgency risk structure contained in continuous spatio-temporal signals; these signals must be transformed into semantic forms that the model can perform reasoning on [Driess *et al.*, 2023; Liang *et al.*, 2022]. Secondly, at the decision-making level, a single LLM simultaneously undertaking high-level decision-making and fine motor execution leads to hierarchical mixing and reasoning conflicts [Song *et al.*, 2023]. Frequent rewriting of local plans by high-level semantic reasoning causes decision jitter and repeated re-planning, and the low-level actions are diluted by high-level language-oriented goals, resulting in outputs that don't satisfy the constraints of the action space. Therefore, separating semantic decision-making from tactical execution is the key to maintaining the stability and executability of LLM reasoning in dynamic social environments [Pateria *et al.*, 2021].

To address the above challenges, we propose AHead-VLN, a human-aware navigation framework that incorporates prospective temporal risks at the language level for decision-making. This integration doesn't simply add additional signals, instead, it transforms short-term occupancy predictions into interpretable and comparable risk alerts, embedding an understanding of future risks into the core of the decision-making process. By doing so, we not only enhance the safety and success rate of intelligent agents in dynamic crowd environments, but also improve the stability and interpretability of the decision-making process. This approach is expected to make navigation behaviors in complex dynamic crowd scenarios more in line with the decision-making logic and social norms of humans [Dong *et al.*, 2025].

The main contributions are summarized as follows: (1) We have designed a path-risk association module (PRA), which is a transition from neural to symbolic compiler. It converts the future spatial-temporal occupancy prediction into a risk description in terms of paths and time indices, providing an interpretable and repeatable interface for LLM planning. (2) We have designed a hierarchical language model planner (HLP), which divides the responsibilities of LLM at the cognitive level: the Brain LLM accomplishes route selection by balancing semantic goals and predicting risks, while the Cerebellum LLM simultaneously generates movement plans for multiple candidate paths. This division method enhances decision consistency by reducing cross-level interference and reduces LLM's latency through parallel processing, thereby improving response capabilities. (3) We conduct extensive experiments and ablation studies on the HA-VLN benchmark, showing consistent improvements over strong reactive and numerical anticipatory baselines in terms of success rate, safety, and decision stability.

2 Related Work

Vision-and-Language Navigation. VLN aims to enable embodied agents to understand natural language instructions and make navigation decisions in visual environments [Anderson *et al.*, 2018; Zhang *et al.*, 2024b]. Early work was mostly conducted in discrete simulation environments based on predefined navigation maps such as Matterport3D[Chang *et al.*, 2017], enhancing the grounding ability of VLN by extending task settings and supervisory signals [Ku *et al.*, 2020]. To bridge the gap between simulation and real-world deployment, [Krantz *et al.*, 2020] introduced the VLN in Continuous Environments (VLN-CE) benchmark, extending VLN from discrete navigation graphs to continuous 3D spaces and thus better reflecting realistic execution processes. Subsequent studies inherit cross-modal alignment strategies from discrete VLN and commonly adopt a two-stage paradigm, in which language-conditioned waypoints are predicted and then executed by a local navigator, connecting traditional VLN with continuous navigation settings [Hong *et al.*, 2022; An *et al.*, 2024]. However, most VLN-CE methods still rely on static assumptions in modeling and evaluation, making them insufficient to handle anticipatory collision risks and personal-space constraints introduced by dynamic pedestrians. HA-VLN [Dong *et al.*, 2025] incorporates safety metrics such as collision rate in addition to the indicators, and examines the prediction of dynamic population changes and the timing strategies of waiting and detouring under occlusion and sensor noise. Thus, the task shifts from static spatial path planning to spatio-temporal risk decision-making for semantic targets.

Social Navigation and Collision Avoidance. Social navigation aims to generate motions that are both safe and socially acceptable in human–robot coexisting environments. Beyond the hard requirement of collision avoidance, it also needs to respect soft social norms such as proxemics, yielding behaviors, and human comfort [Singamaneni *et al.*, 2024; Gao and Huang, 2022]. To address the short-sightedness of classical reactive strategies in complex interactions, existing studies mainly advance along three directions. Firstly, they combine prediction with planning by mapping the trajectories or occupancy predictions onto time-varying cost maps or risk areas, in order to achieve forward reasoning [Poddar *et al.*, 2023; Liu *et al.*, 2022; Le *et al.*, 2024; Salzmann *et al.*, 2020;

Yuan *et al.*, 2021]. Secondly, they leverage deep reinforcement learning to incorporate social costs, risk regions, or action constraints, balancing efficiency and compliance [Xue *et al.*, 2024; Feng *et al.*, 2024; Cheng *et al.*, 2024; Flögel *et al.*, 2024]. Thirdly, they integrate uncertainty awareness and online risk adaptation, adjusting risk sensitivity based on confidence to mitigate the robustness issues caused by prediction errors [Sun *et al.*, 2025]. Although significant progress has been made, most methods still rely on numerical costs, which makes it difficult to match spatial-temporal risks with high-level intentions in language-influenced environments. This has prompted us to adopt a more structured risk representation approach for decision-making.

Large Language Models for VLN. In recent years, large language models (LLMs) have been incorporated into the advanced decision-making of visual and language navigation (VLN), achieving route selection and action planning through language-based reasoning [Zhang *et al.*, 2024b]. Existing methods typically connect text reasoning with executable navigation in different ways. For instance, SayCan selects skills or actions based on feasibility scores to ensure executability [Brohan *et al.*, 2023]. LLM-Planner generates structured intermediate plans and implements them through low-level modules, thereby reducing illusions and unfeasible outputs [Song *et al.*, 2023]. ReAct improves the robustness of online decision making via a cyclic interaction of reasoning, action, and feedback [Yao *et al.*, 2022]. In contrast, NavGPT [Zhou *et al.*, 2024], DiscussNav [Sun *et al.*, 2025], and MapGPT [Chen *et al.*, 2024] place greater emphasis on organizing observations into prompts, performing explicit reasoning, and outputting decisions at the action or path level. However, under the social awareness constraints defined by HA-VLN, navigation needs to predict human actions while meeting the requirements for collision avoidance and personal space. In this case, relying solely on language reasoning is insufficient to support stable decisions. Without structured input that clearly indicates the dynamics of the crowd and personal space, language models usually struggle to reliably generate time-sensitive strategies such as waiting and detouring. Therefore, we have designed the HLP, which separates high-level language reasoning from continuous control and risk handling to enhance the stability and reliability of navigation decisions in dynamic and populated environments.

3 Method

Task Setup. In this work, we address Vision-and-Language Navigation in Continuous Environments (VLN-CE). The agent navigates within a continuous 3D space $E \in \mathbb{R}^3$, where its pose at time step t is denoted by $X_t = (x_t, y_t, z_t) \in E$, representing its absolute coordinate. To perceive the environment, the agent captures panoramic RGB-D observations from a set of discrete, uniformly distributed viewpoints $\Theta = 0°, 30°, \ldots, 330°$. This yields a set of 12 RGB and 12 depth images, denoted as $I_t = (I_{t,i}^{rgb}, I_{t,i}^{depth}) \mid i = 1, \ldots, 12$, where $I_{t,i}^{rgb} \in \mathbb{R}^{H \times W \times 3}$and $I_{t,i}^{depth} \in \mathbb{R}^{H \times W}$. The agent is provided with a natural language instruction composed of a sequence of tokens $L = l_1, l_2, \ldots, l_M$, describing the intended path from a starting pose $X_{start} \in E$ to a target location $X_{goal} \in E$. Navigation is performed within a continuous 3D mesh environment using a discrete low-level action space $\mathcal{A}$ (e.g., move forward, turn left, turn right). The objective of this agent is to reach the target position X_{goal} while adhering to the instruction L, thereby minimizing the final navigation error.

Overview of our Approach. We propose AHead-VLN to address a key challenge in human-crowded, dynamic environments: VLN decision-making lacks anticipatory risk understanding and therefore struggles with semantic reasoning and real-time planning. As illustrated in Figure 2, our method takes the current panoramic RGB-D observation I_t and the language instruction L as input, and produces an executable low-level action sequence. Overall, the pipeline of AHead-VLN goes from perception to compilation to planning. First, a waypoint predictor $f_\theta^{waypoint}$ generates a set of candidate paths $\mathcal{P} = \{P_k\}_{k=1}^K$, while a spatio-temporal occupancy forecasting module $f\phi^{pred}$ produces a sequence of high-dimensional future spatio-temporal occupancy grid maps $\mathcal{M}_{SOGM}$ [Agro *et al.*, 2023]. The former provides geometrically feasible options consistent with the instruction, and the latter captures the anticipated evolution of static, movable, and dynamic entities over the next few seconds. Additionally, we introduce PRA module f_ψ^{PRA}, aligning the continuous spatio-temporal risks with the discrete path key points, and compiling the high-dimensional risk information into a tactical summary S_k for global strategy selection, as well as a precise risk representation of conflict event time and location R_k. This enables the planner to carry out time-aware actions such as waiting or yielding. Finally, decision-making is carried out by the HLP. The high-level Brain reasons over $\{S_k\}_{k=1}^K$ and selects the best path index k^*, while the low-level Cerebellum generates an executable action script Π_k for each candidate path in parallel. Once the Brain makes its choice, the system directly retrieves Π_{k^*} as the final plan, which reduces decision latency via parallel generation and fast retrieval, and improves behavioral stability while maintaining semantic reasoning capability.

3.1 Dual-Stream Perception

Waypoint Prediction. The main objective of this approach is to generate geometrically feasible candidate paths. We use the waypoint predictor $f_\theta^{waypoint}$ to generate candidate waypoints based on the current observations, bridging the gap between the discrete environment and the continuous space [Hong *et al.*, 2022]. The predictor consists of three components. First, Visual Encoders (ResNet-50) extract features from the current panoramic RGB-D observation I_t. Second, a spatial relation transformer models the global scene layout by integrating the features from 12 discrete perspectives. Finally, the waypoint heat map predictor outputs the probability distribution of navigable points and uses non-maximum suppression (NMS) to extract K discrete candidate waypoints $\{ w_k \mid k = 1, \ldots, K \}$. For each waypoint w_k, the local path planner generates a kinematically feasible trajectory from the current position and downsamples it to a sequence of key points with a length of L, denoted as

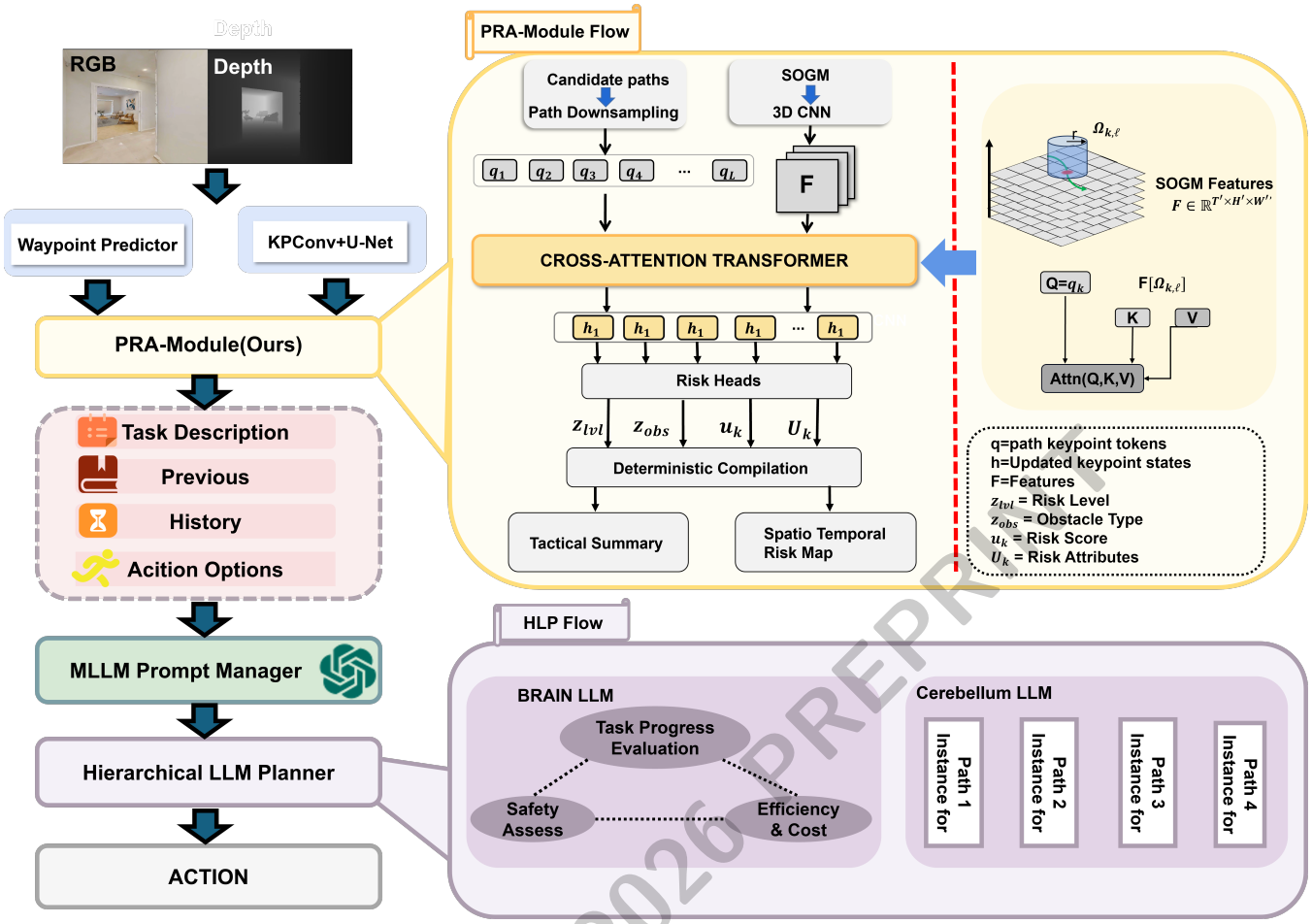

Figure 2: The AHead-VLN consists of two key components: the Path–Risk Association (PRA) module and the navigation decision module based on the hierarchical language model (HLP). The PRA module, by sensing the candidate paths and future space occupancy situations, uses a path-based sparse cross-attention Transformer to compile future risks into a structured representation, thereby enhancing the interpretability and usability of risk understanding. In the HLP, the Brain selects the optimal path based on the summary, while the Cerebellum generates the corresponding action script in parallel, aiming to reduce the planning delay while maintaining the quality of decision-making.

$\{\mathcal{P} = P_k \mid k = 1, \ldots, K\}$. These paths $\mathcal{P}$ serve as the geometric queries for the subsequent risk association module.

Proactive Spatio-temporal Forecasting. Parallel to the generation of waypoints, the proactive spatio-temporal prediction also operates independently. To capture comprehensive scene dynamics, we synthesize the panoramic depth observations into an approximately 360° point cloud C_t, extract point-wise semantic features using a KPConv-based 3D U-Net, and project them onto a bird's-eye-view (BEV) feature map. A 2D front-end with temporal propagation blocks then performs sequential forecasting on the BEV representation to produce future occupancy grids; we further fine-tune this network in the HA-VLN environment [Thomas *et al.*, 2023]. The network outputs a sequence of future spatio-temporal occupancy grid maps (SOGM), denoted as $\mathcal{M}_{SOGM} \in \mathbb{R}^{C\times T\times H\times W}$, where $H \times W$ is the BEV resolution and T is the prediction horizon. Each channel encodes the occupancy probability of a different obstacle source (e.g., static, movable, and dynamic), enabling source-aware risk assessment.

3.2 Neural-to-Symbolic Compilation

To bridge the gap between raw perception data and LLM-based planners, we introduce the PRA module. The PRA module compiles SOGM into compact, structured risk representations aligned with candidate paths. These representations serve as tactical summaries and spatio-temporal risk annotations for hierarchical LLM reasoning. The module operates in two stages: path-conditioned symbol grounding and risk schema compilation.

Path-Guided Spatio-temporal Grounding. The raw $\mathcal{M}_{SOGM}$ is high-dimensional and sparse. We employ a lightweight 3D convolutional encoder, Φenc to compress the $\mathcal{M}_{SOGM}$ into a spatio-temporal feature embedding F:

$$F = \Phi_{\text{enc}}(\mathcal{M}_{\text{SOGM}}) \in \mathbb{R}^{T'\times H'\times W'\times d} \quad (1)$$

To geometrically align risk features with each candidate path, we resample every path P_k into a fixed-length keypoint sequence $P_k = \{p_{k,\ell}\}_{\ell=1}^{L}$, and build a path-environment interaction module by stacking M interaction layers. Each

layer follows a standard Transformer decoder structure. We focus on the proposed path-guided sparse cross-attention: for each keypoint $p_{k,\ell}$, we define a local spatio-temporal neighborhood centered at its projected location on the feature grid, which can be approximated as a "virtual cylinder" with radius r . Let $\Omega_{k,\ell}$ denote the sparse sampling index set within this neighborhood.

$$\Omega_{k,\ell} = \{(u,v) \mid \|(u,v) - \text{proj}(p_{k,\ell})\|_2 \le r\}, \quad (2)$$

where $\text{proj}(\cdot)$ maps world coordinates to feature grid indices. We then perform sparse multi-head cross-attention, where the query $q_{k,\ell}$ interacts only with keys/values extracted from $\Omega_{k,\ell}$. For the j-th attention head, the sparse multi-head cross-attention is computed as:

$$\text{Head}_j(q_{k,\ell}, \Omega_{k,\ell}) = \text{Softmax}\left(\frac{q_{k,\ell} W_Q^j (K_{\Omega_{k,\ell}} W_K^j)^T}{\sqrt{d_k}}\right) \cdot (V_{\Omega_{k,\ell}} W_V^j). \quad (3)$$

This mechanism enforces a physical locality constraint by restricting cross-attention to regions consistent with the path geometry, leading to more geometrically grounded risk attribution.After M iterative interactions, we obtain the risk-aware hidden-state sequence $\{h_\ell^k\}_{\ell=1}^L$, where h_ℓ^k is the hidden token (decoder hidden state) of the ℓ-th keypoint, and serves as the input to subsequent risk prediction and structured compilation.

Risk Schema Compilation. Based on the keypoint representations $\{h_\ell^k\}_{\ell=1}^L$, the PRA module predicts a concise path-level risk summary as well as the spatio-temporal risks along the path, and deterministically compiles it into a structured risk schema that can be used by the LLM planner. Specifically, the tactical summary head outputs $z_{\text{lvl}}^k \in \mathbb{R}^4$ and $z_{\text{obs}}^k \in \mathbb{R}^3$, while the spatio-temporal risk head outputs $u^k \in \mathbb{R}^L$ and $U^k \in \mathbb{R}^{L\times 8}$. To ensure the stability and repeatability of the interface, we map the logarithm of the risk level to discrete labels in the following way:

$$p_{\text{lvl}}^k = \text{softmax}(z_{\text{lvl}}^k), \text{level}^k = \arg\max p_{\text{lvl}}^k. \quad (4)$$

The observation feature z_{obs}^k is mapped to obstacle-source indicators o^k. The tactical summary is then defined as $S_k = (\text{level}^k, \text{o}^k)$. Next, threshold-based segmentation and interval merging are applied to (u^k, U^k) to obtain a set of risk events $R_k = \{e_j\}_{j=1}^{N_k}$, where each event is represented as

$$e = (\text{id}, \mathcal{T}, t_{\text{start}}, t_{\text{end}}, s_{\text{start}}, s_{\text{end}}, \text{description}), \quad (5)$$

with id being a unique event identifier within path k. $t_{\text{start}}, t_{\text{end}}$ denote the predicted time window (as discrete forecast-step indices, convertible to seconds via Δt), and $s_{\text{start}}, s_{\text{end}}$ denote the affected path segment. The description is a short templated natural-language phrase used in the LLM prompt. The $\mathcal{T}$ is a discrete label derived from level^k and the primary obstacle source inferred from z_{obs}^k.

In this final result, the Brain component uses S_k to compare candidate paths, while the Cerebellum component uses R_k to avoid high-risk intervals when generating action scripts explicitly. This enables the hierarchical LLM planning to maintain controllability and interpretability without directly exposing high-dimensional spatial tensors.

3.3 Hierarchical LLM Planner

After receiving the compiled symbolic outputs from the PRA module, the system enters the final decision phase. We propose a Parallel Hierarchical LLM Planner (HLP) that decomposes planning into two stages: (i) strategic option selection over candidate paths and (ii) tactical action scripting conditioned on the selected path and its associated risk descriptors. The overall prompt structure and the parallel two-stage workflow are illustrated in Figure 3.

The Brain: Strategic Option Selection. As a high-level planner, the Brain LLM is not responsible for generating low-level action sequences. Instead, it evaluates trade-offs and selects a preferred option among K candidate paths. Concretely, we formulate this step as a multimodal option selection problem driven by a structured prompt.

The global context provides essential background for decision-making. It includes the task instruction $\mathcal{L}$, the previous plan P_{t-1} summarizing recently executed steps, and a condensed history H_t (with a fixed token budget) that summarizes visited locations and key observations. The second part of the prompt is the candidate set. For each path $k \in \{1, \ldots, K\}$, the system constructs an option described by three elements: a directional description D_k, a corresponding viewpoint image v_k, and a tactical summary S_k produced by the PRA module. The Brain policy π_{Brain} integrates the above information and outputs an index k^* indicating the selected option:

$$k^* = \pi_{\text{Brain}}\left(\mathcal{L}, P_{t-1}, H_t, \{(D_k, v_k, S_k)\}_{k=1}^K\right). \quad (6)$$

This design conditions option selection on both task requirements and risk cues, aiming to favor decisions that are better aligned with the instruction while reducing predicted risk exposure.

The Cerebellum: Parallel Tactical Grounding. In parallel with the Brain's deliberation, a Cerebellum cluster performs symbol-to-action scripting. Unlike the Brain, the Cerebellum focuses on how to execute a given candidate path segment under the provided risk descriptors. For each candidate path P_k, we instantiate a dedicated Cerebellum module $\pi_{\text{Cereb}}^{(k)}$. It receives (i) the assigned local path segment $\widetilde{\mathcal{P}}_k$ (e.g., a waypoint sequence for the current horizon), (ii) detailed spatio-temporal risk information R_k generated by the PRA module, and (iii) a predefined discrete action space $\mathcal{A}$. The Cerebellum then produces an action script Π_k composed of primitives from $\mathcal{A}$:

$$\Pi_k = \pi_{\text{Cereb}}^{(k)}\left(\widetilde{\mathcal{P}}_k, R_k, \mathcal{A}\right), \forall a \in \Pi_k \cap \mathcal{A}. \quad (7)$$

Compared with purely reactive policies, conditioning the script on predicted spatio-temporal risk descriptors enables anticipatory maneuvers (e.g., waiting or detouring) rather than responding only to immediate observations. Finally, we cache the output results of the Cerebellum as the candidate script pool. Then, the Execution Selector will match the decision $k^*$of the brain with the corresponding script and obtain Π_{k^*} for execution. Since this selection is a lightweight lookup, the overall planner reduces latency while maintaining controllability via decoupled reasoning and scripting.

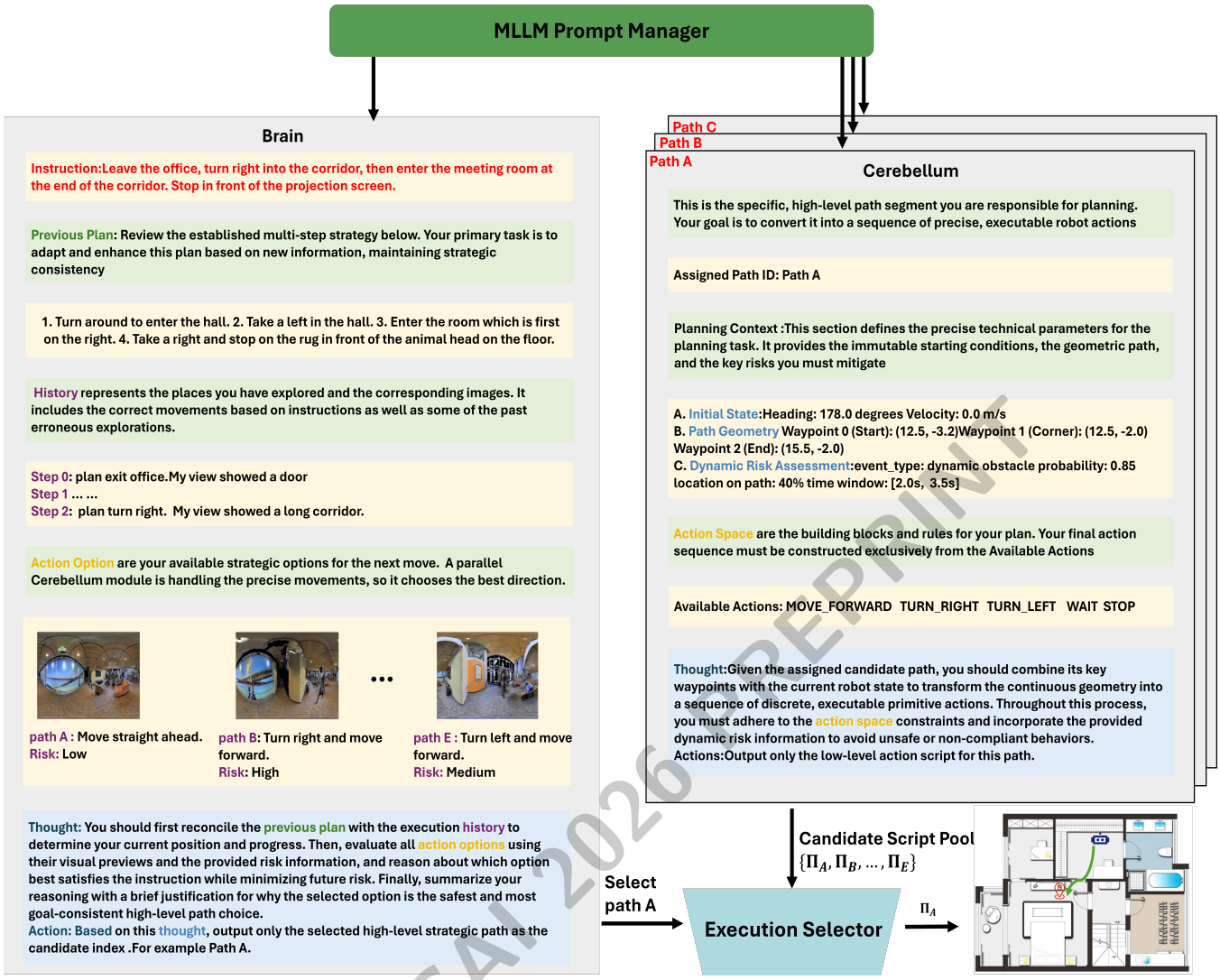

Figure 3: Given the instructions, the prompt manager first organizes the previous plans and explores the history, then presents the path options with risk markers to the brain for selection of the next path. Meanwhile, it provides the risk events and action space for each selected path to the cerebellum for reasoning to achieve action planning. Finally, the executor uses the option selected by the Brain to output the matching action sequence for execution, thereby enabling rapid switching when the decision changes.

4 Experiment

4.1 Experiment Setup

Simulated Environments. Our implementation is based on the HA-VLN simulator and is benchmarked on the HA-R2R instruction dataset. This dataset contains a total of 16,844 natural language instructions, covering 90 building scans. Additionally, HA-VLN provides human-related dynamic configurations (910 human models, 428 regions) in the scenes of the 90 scans, thus making the navigation evaluation more closely resemble real indoor scenarios. The agent's input is language instructions and panoramic visual observations (RGB-D), and the output is a sequence of discrete actions. To assess zero-shot generalization across different scenes, we use a sampled subset of 90 scenarios and 216 trajectories for evaluation. We conducted comparative experiments with a range of learning-based VLN methods and LLM/multi-LLM-based approaches to evaluate the navigator's performance. Among them, the PRA module is a learning-based module. During the training process, we used the AdamW optimizer, with a learning rate of 10^{-4} and batch size of 4 on two NVIDIA A100 GPUs. In the zero-shot inference stage, the planning module is deployed through the API: Brain uses gpt-4o-2024-05-13, and Cerebellum uses gpt-3.5-turbo.

Evaluation Metrics. We use the HA-VLN evaluation metrics to assess the agent: Navigation Error (NE) represents the average Euclidean distance between the agent's position at the end of navigation and the target point, which is used to measure the arrival accuracy. The total collision rate (TCR) counts the number of collisions between the agent and pedestrians in each episode and takes the average to reflect the overall collision intensity. TCR quantifies the frequency of colli-

sions that occur in areas with human activity. The collision rate (CR) calculates the proportion of episodes in which at least one collision occurs, reflecting the probability of collision occurrence. Finally, the success rate (SR) counts the agent's success rate in the HA-VLN environment when it is within 3 meters of the target and no collision occurs throughout the process. These standards ensure that while taking into account the navigation accuracy, they can objectively measure its safety and compliance with spatial constraints in dynamic crowd scenarios.

4.2 Results in Simulating Environments

Agent	Setting	NE ↓	TCR ↓	CR ↓	SR ↑
HA-VLN-CMA-Base	Retrained	7.34	47.06	0.77	0.07
	Zero-shot	7.95	63.96	0.76	0.03
HA-VLN-CMA-DA	Retrained	7.00	27.25	0.69	0.09
	Zero-shot	7.05	38.22	0.73	0.05
HA-VLN-CMA*	Retrained	6.23	8.10	0.69	0.10
	Zero-shot	6.62	32.48	0.70	0.09
HA-VLN-VL	Retrained	5.35	6.63	0.59	0.14
	Zero-shot	7.15	3.97	0.46	0.06
ETPNav	Retrained	5.51	4.71	0.55	0.21
	Zero-shot	6.10	5.72	0.56	0.15
BEVBert	Retrained	5.43	6.94	0.58	0.17
	Zero-shot	7.40	7.94	0.71	0.08
AHead-VLN (Ours)	**Zero-shot**	5.68	**3.89**	**0.44**	**0.17**

Table 1: COMPARISON ON SIMULATED ENVIRONMENT IN HA-R2R DATASET

Table 1 presents the performance of different methods on HA-R2R under the Retrained and Zero-shot settings. Retrained refers to the strong supervision assumption of further training on the target environment, which usually significantly improves the success rate or reduces the error; while Zero-shot is closer to real deployment, requiring the model to generalize and make safe decisions without additional training. By comparison, many traditional methods often rely on Retraining to achieve better performance, and their Zero-shot performance generally declines. In contrast, AHead-VLN still maintains a high safety rate and success rate even under a complete Zero-shot condition. In the zero-shot setting, our agent achieves an SR of 0.17, which is the best result among the compared methods; relative to the baseline with the highest SR (0.15), this corresponds to an improvement of approximately 13.3%. Since SR in HA-VLN requires not only reaching the target (within 3 meters) but also avoiding collisions throughout the entire episode, this gain demonstrates the simultaneous satisfaction of task completion and safety constraints. From the safety perspective, AHead-VLN achieves TCR and CR, indicating both lower average collision intensity and lower collision occurrence probability; compared with the safest baseline, we further reduce TCR and CR by about 2.0% and 4.35%, respectively. This suggests that the proposed method can perceive and avoid risks more effectively without training-time adaptation, thereby improving the reliability and safety of navigation. Meanwhile, NE remains within a reasonable range, increasing by only 2.46% compared with the best NE baseline, which indicates that the safety benefits do not come from excessive conservatism or a substantial loss of efficiency. Overall, AHead-VLN achieves more reliable safety and stronger collision-free success without retraining.

4.3 Ablation Study

Variant	NE ↓	TCR ↓	CR ↓	SR ↑
Full AHead-VLN	**5.68**	**3.89**	**0.44**	**0.17**
(1) w/o PRA	7.53	23.12	0.74	0.06
(2) Serial scripts	5.79	5.76	0.56	0.12
(3) w/o Cerebellum (Controller)	5.83	5.41	0.53	0.14
(4) w/o PRA + w/o Cerebellum	7.61	25.48	0.76	0.02

Table 2: ABLATION STUDY ON PRA AND HLP

Table 2 shows the ablation results, indicating that each module contributes to the safe arrival. Firstly, PRA is the core of safety: after its removal, the increases in TCR and CR were significant, while SR decreased from 0.17 to 0.06, and NE also deteriorated, indicating that the lack of structured forward risk information would simultaneously undermine risk avoidance and goal advancement. Secondly, changing the parallel script to a serial one would cause TCR and CR to rise to 5.76/0.56, SR to drop to 0.12, and NE to change only slightly, suggesting that the parallel mechanism mainly improves the decision-making timeliness and process stability in dynamic scenarios, rather than the static navigation accuracy. Again, after removing Cerebellum, TCR became 5.41, CR became 0.53, and SR dropped to 0.14, while NE only slightly increased. This indicates that hierarchical control is crucial for the controllability of action execution and fine-grained avoidance. Finally, when both PRA and Cerebellum are removed simultaneously, the result is the worst, verifying their complementarity: PRA determines "where / when to avoid", and HLP ensures "how to safely and stably execute".

The ablation results, together with qualitative failure cases, suggest two remaining limitations: PRA may still misestimate localized risks in occluded layouts such as corners or pillars, while zero-shot LLM planning and Cerebellum scripting may suffer from ambiguous instruction grounding and imperfect action timing

5 Conclusion

We propose AHead-VLN, a human-aware navigation framework that converts predicted risks into path-level event descriptions for hierarchical LLM planning. By separating strategic path selection from tactical action execution, AHead-VLN improves decision stability, interpretability, and safety in dynamic crowd environments. Experiments on HA-VLN demonstrate better zero-shot success and collision avoidance than strong baselines. Future work will explore broader VLN scenarios and real-world deployment.

Acknowledgements

This paper is supported by the Major Scientific and Technological Research Project of Hebei Communications Investment Group Co., Ltd., "Research and Application of BeiDou + 5G Positioning and Navigation Technology" under grant KT2024D1-01.

This paper is supported by the Hebei Provincial Major Science and Technology Support Program project "Beidou + Indoor Distributed Positioning and Navigation Application Demonstration for the Large-Scale Underground Space in Xiong'an New Area" under grant 242Q040Z.

References

[Agro *et al.*, 2023] Ben Agro, Quinlan Sykora, Sergio Casas, and Raquel Urtasun. Implicit occupancy flow fields for perception and prediction in self-driving. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1379–1388, 2023.

[An *et al.*, 2024] Dong An, Hanqing Wang, Wenguan Wang, Zun Wang, Yan Huang, Keji He, and Liang Wang. ETPNav: Evolving Topological Planning for Vision-Language Navigation in Continuous Environments. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–16, 2024.

[Anderson *et al.*, 2018] Peter Anderson, Qi Wu, Damien Teney, Jake Bruce, Mark Johnson, Niko Sünderhauf, Ian Reid, Stephen Gould, and Anton van den Hengel. Vision-and-language navigation: Interpreting visually-grounded navigation instructions in real environments. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[Biswas *et al.*, 2022] Abhijat Biswas, Allan Wang, Gustavo Silvera, Aaron Steinfeld, and Henny Admoni. Socnavbench: A grounded simulation testing framework for evaluating social navigation. *ACM Transactions on Human-Robot Interaction (THRI)*, 11(3):1–24, 2022.

[Brohan *et al.*, 2023] Anthony Brohan, Yevgen Chebotar, Chelsea Finn, Karol Hausman, Alexander Herzog, Daniel Ho, Julian Ibarz, Alex Irpan, Eric Jang, Ryan Julian, et al. Do as i can, not as i say: Grounding language in robotic affordances. In *Conference on robot learning*, pages 287–318. PMLR, 2023.

[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. *arXiv preprint arXiv:1709.06158*, 2017.

[Chen *et al.*, 2024] Jiaqi Chen, Bingqian Lin, Ran Xu, Zhenhua Chai, Xiaodan Liang, and Kwan-Yee Wong. Mapgpt: Map-guided prompting with adaptive path planning for vision-and-language navigation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9796–9810, 2024.

[Cheng *et al.*, 2024] Chien-Lun Cheng, Chen-Chien Hsu, Saeed Saeedvand, and Jun-Hyung Jo. Multi-objective crowd-aware robot navigation system using deep reinforcement learning. *Applied Soft Computing*, 151:111154, 2024.

[Dong *et al.*, 2025] Yifei Dong, Fengyi Wu, Qi He, Heng Li, Minghan Li, Zebang Cheng, Yuxuan Zhou, Jingdong Sun, Qi Dai, Zhi-Qi Cheng, and Alexander G. Hauptmann. HA-VLN: A Benchmark for Human-Aware Navigation in Discrete-Continuous Environments with Dynamic Multi-Human Interactions, Real-World Validation, and an Open Leaderboard, May 2025. arXiv:2503.14229 [cs].

[Driess *et al.*, 2023] Danny Driess, Fei Xia, Mehdi SM Sajjadi, Corey Lynch, Aakanksha Chowdhery, Ayzaan Wahid, Jonathan Tompson, Quan Vuong, Tianhe Yu, Wenlong Huang, et al. Palm-e: An embodied multimodal language model. 2023.

[Feng *et al.*, 2024] Zhen Feng, Bingxin Xue, Chaoqun Wang, and Fengyu Zhou. Safe and socially compliant robot navigation in crowds with fast-moving pedestrians via deep reinforcement learning. *Robotica*, 42(4):1212–1230, 2024.

[Flögel *et al.*, 2024] Daniel Flögel, Lars Fischer, Thomas Rudolf, Tobias Schürmann, and Sören Hohmann. Socially integrated navigation: A social acting robot with deep reinforcement learning. In *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 4785–4792. IEEE, 2024.

[Gao and Huang, 2022] Yuxiang Gao and Chien-Ming Huang. Evaluation of socially-aware robot navigation. *Frontiers in Robotics and AI*, 8:721317, 2022.

[Gu *et al.*, 2022] Jing Gu, Eliana Stefani, Qi Wu, Jesse Thomason, and Xin Wang. Vision-and-language navigation: A survey of tasks, methods, and future directions. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7606–7623, 2022.

[Hong *et al.*, 2022] Yicong Hong, Zun Wang, Qi Wu, and Stephen Gould. Bridging the Gap Between Learning in Discrete and Continuous Environments for Vision-and-Language Navigation. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15418–15428, New Orleans, LA, USA, June 2022. IEEE.

[Krantz *et al.*, 2020] Jacob Krantz, Erik Wijmans, Arjun Majumdar, Dhruv Batra, and Stefan Lee. Beyond the Nav-Graph: Vision-and-Language Navigation in Continuous Environments. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, editors, *Computer Vision – ECCV 2020*, pages 104–120, Cham, 2020. Springer International Publishing.

[Ku *et al.*, 2020] Alexander Ku, Peter Anderson, Roma Patel, Eugene Ie, and Jason Baldridge. Room-Across-Room: Multilingual Vision-and-Language Navigation with Dense Spatiotemporal Grounding. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu, editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4392–4412, Online,

November 2020. Association for Computational Linguistics.

[Le *et al.*, 2024] Viet-Anh Le, Behdad Chalaki, Vaishnav Tadiparthi, Hossein Nourkhiz Mahjoub, Jovin D'sa, and Ehsan Moradi-Pari. Social navigation in crowded environments with model predictive control and deep learning-based human trajectory prediction. In *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 4793–4799. IEEE, 2024.

[Liang *et al.*, 2022] Jacky Liang, Wenlong Huang, Fei Xia, Peng Xu, Karol Hausman, Brian Ichter, Pete Florence, and Andy Zeng. Code as policies: Language model programs for embodied control. *arXiv preprint arXiv:2209.07753*, 2022.

[Liu *et al.*, 2022] Shuijing Liu, Peixin Chang, Zhe Huang, Neeloy Chakraborty, Kaiwen Hong, Weihang Liang, D Livingston McPherson, Junyi Geng, and Katherine Driggs-Campbell. Intention aware robot crowd navigation with attention-based interaction graph. *arXiv preprint arXiv:2203.01821*, 2022.

[Möller *et al.*, 2021] Ronja Möller, Antonino Furnari, Sebastiano Battiato, Aki Härmä, and Giovanni Maria Farinella. A survey on human-aware robot navigation. *Robotics and Autonomous Systems*, 145:103837, 2021.

[Paez-Granados *et al.*, 2022] Diego Paez-Granados, Vaibhav Gupta, and Aude Billard. Unfreezing social navigation: Dynamical systems based compliance for contact control in robot navigation. In *2022 International Conference on Robotics and Automation (ICRA)*, pages 8368–8374. IEEE, 2022.

[Park and Kim, 2023] Sang-Min Park and Young-Gab Kim. Visual language navigation: A survey and open challenges. *Artificial Intelligence Review*, 56(1):365–427, 2023.

[Pateria *et al.*, 2021] Shubham Pateria, Budhitama Subagdja, Ah-hwee Tan, and Chai Quek. Hierarchical reinforcement learning: A comprehensive survey. *ACM Computing Surveys (CSUR)*, 54(5):1–35, 2021.

[Poddar *et al.*, 2023] Sriyash Poddar, Christoforos Mavrogiannis, and Siddhartha S Srinivasa. From crowd motion prediction to robot navigation in crowds. In *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 6765–6772. IEEE, 2023.

[Salzmann *et al.*, 2020] Tim Salzmann, Boris Ivanovic, Punarjay Chakravarty, and Marco Pavone. Trajectron++: Dynamically-feasible trajectory forecasting with heterogeneous data. In *European Conference on Computer Vision*, pages 683–700. Springer, 2020.

[Sathyamoorthy *et al.*, 2020] Adarsh Jagan Sathyamoorthy, Utsav Patel, Tianrui Guan, and Dinesh Manocha. Frozone: Freezing-free, pedestrian-friendly navigation in human crowds. *IEEE Robotics and Automation Letters*, 5(3):4352–4359, 2020.

[Singamaneni *et al.*, 2024] Phani Teja Singamaneni, Pilar Bachiller-Burgos, Luis J Manso, Anaís Garrell, Alberto Sanfeliu, Anne Spalanzani, and Rachid Alami. A survey on socially aware robot navigation: Taxonomy and future challenges. *The International Journal of Robotics Research*, 43(10):1533–1572, 2024.

[Song *et al.*, 2023] Chan Hee Song, Jiaman Wu, Clayton Washington, Brian M Sadler, Wei-Lun Chao, and Yu Su. Llm-planner: Few-shot grounded planning for embodied agents with large language models. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 2998–3009, 2023.

[Sun *et al.*, 2025] Yanjun Sun, Yue Qiu, and Yoshimitsu Aoki. DynamicVLN: Incorporating Dynamics into Vision-and-Language Navigation Scenarios. *Sensors*, 25(2):364, January 2025. Number: 2 Publisher: Multidisciplinary Digital Publishing Institute.

[Thomas *et al.*, 2023] Hugues Thomas, Jian Zhang, and Timothy D Barfoot. The foreseeable future: Self-supervised learning to predict dynamic scenes for indoor navigation. *IEEE Transactions on Robotics*, 39(6):4581–4599, 2023.

[Xue *et al.*, 2024] Bingxin Xue, Ming Gao, Chaoqun Wang, Yao Cheng, and Fengyu Zhou. Crowd-aware socially compliant robot navigation via deep reinforcement learning. *International Journal of Social Robotics*, 16(1):197–209, 2024.

[Yao *et al.*, 2022] Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik R Narasimhan, and Yuan Cao. React: Synergizing reasoning and acting in language models. In *The eleventh international conference on learning representations*, 2022.

[Yuan *et al.*, 2021] Ye Yuan, Xinshuo Weng, Yanglan Ou, and Kris M Kitani. Agentformer: Agent-aware transformers for socio-temporal multi-agent forecasting. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9813–9823, 2021.

[Zhang *et al.*, 2024a] Yue Zhang, Ziqiao Ma, Jialu Li, Yanyuan Qiao, Zun Wang, Joyce Chai, Qi Wu, Mohit Bansal, and Parisa Kordjamshidi. Vision-and-language navigation today and tomorrow: A survey in the era of foundation models. *arXiv preprint arXiv:2407.07035*, 2024.

[Zhang *et al.*, 2024b] Yue Zhang, Ziqiao Ma, Jialu Li, Yanyuan Qiao, Zun Wang, Joyce Chai, Qi Wu, Mohit Bansal, and Parisa Kordjamshidi. Vision-and-Language Navigation Today and Tomorrow: A Survey in the Era of Foundation Models, July 2024. arXiv:2407.07035 [cs] version: 1.

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