HCSG: Human-Centric Semantic-Geometric Reasoning for Vision-Language Navigation

Haoxuan Xu, Tianfu Li, Wenbo Chen, Yi Liu, Jin Wu, Huashuo Lei, Yunfan Lou, Lujia Wang, Hesheng Wang, Haoang Li

***Abstract*—VLN has achieved remarkable progress by scaling data and model capacity. However, the assumption of a static environment breaks down in real-world indoor scenarios, where robots inevitably encounter dynamic pedestrians. Existing human-aware approaches typically treat humans merely as moving obstacles based on implicit visual cues, lacking the explicit reasoning required to interpret human intentions or maintain social norms. To address this, we propose HCSG, the first human-centric framework for VLN. This framework provides a robust foundation for safe, socially intelligent navigation in dynamic human-robot environments that shifts the paradigm from passive collision avoidance to active human behavior understanding. Specifically, HCSG introduces a unified Human Understanding Module that synergizes two key capabilities: (i) geometric forecasting, which predicts human pose and trajectory to anticipate future motion dynamics; and (ii) semantic interpretation, which leverages a Vision-Language Model (VLM) to generate natural language descriptions of human actions and intentions. These semantic-geometric representations are fused into the agent's topological map for instruction-conditioned planning. Furthermore, a social distance loss is introduced to enforce socially compliant interaction distances. Extensive experiments on the HA-VLNCE benchmark demonstrate that HCSG significantly outperforms state-of-the-art methods, achieving a 14% improvement in Success Rate and a 34% reduction in Collision Rate.**

***Index Terms*—Vision-Language Navigation, Human-aware Understanding, Social Navigation.**

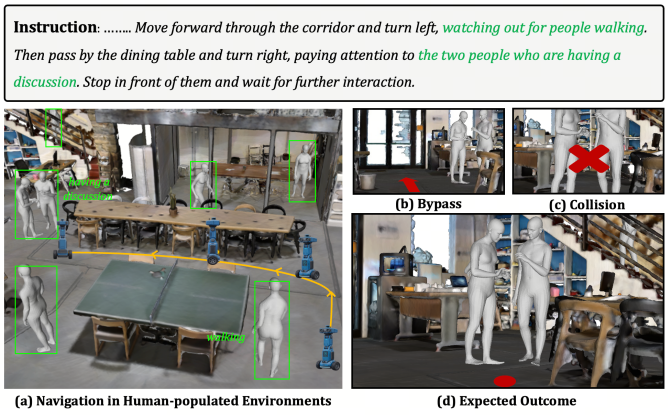


Fig. 1. An illustrative navigation scenario in a human-populated environment. The input is an instruction "*pay attention to the two people who are having a discussion. Stop in front of them...*". (a) **Overview**: A typical scenario demonstrating the challenge of dynamic humans. (b) **Bypass**: Without semantic understanding, traditional agents may misclassify task-relevant people as generic obstacles and avoid them entirely. (c) **Collision**: Without geometric understanding, traditional agents often fail to anticipate future motion, leading to physical collisions. (d) **Expected Outcome**: Our HCSG synergizes semantic interpretation and geometric forecasting to achieve instruction-grounded and socially compliant navigation.

I. Introduction

Vision-Language Navigation (VLN) [1]–[4] enables robots to follow multimodal instructions and navigate physical spaces. Given a language instruction, an embodied agent must interpret commands together with egocentric visual observations to reach a target location or object. Recent advances have propelled VLN from simplified discrete settings [4] to complex continuous environments [2], [5], increasingly leveraging large-scale pre-training and foundation models to improve instruction grounding and embodied reasoning [6]–[8]. As such, VLN has become a key capability for service robots in indoor environments.

However, the prevailing VLN paradigm largely assumes a static environment [3], which breaks down in real-world settings where robots inevitably encounter moving pedestrians. Safe navigation requires anticipating future pedestrian motion, a problem widely studied in trajectory prediction [9]. To narrow this realism gap, recent works have augmented simulators with human-populated scenes [5], [10]–[13]. Yet existing VLN agents still treat pedestrians primarily as moving obstacles rather than task-relevant entities. Their representations rely mainly on global visual cues and lack explicit modeling of human activities, intentions, and social interaction signals. Although several pioneering efforts have introduced partial forms of human awareness, they either restrict interaction to explicit dialogue [14], [15] or use vision-language models mainly for passive collision avoidance [16], [17]. As a result, current navigation policies lack explicit reasoning about what humans are doing, where they are moving, and how the robot should behave around them under social norms. This limitation leads to severe failures in dynamic human-populated environments, as illustrated in Fig. 1. Consider the instruction *"pay attention to the two people... having a discussion and*

Haoxuan Xu, Tianfu Li, Wenbo Chen, Huashuo Lei, Lujia Wang and Haoang Li are with The Hong Kong University of Science and Technology (Guangzhou), Guangzhou 511453, China (e-mail: hxu095@connect.hkust-gz.edu.cn; tli794@connect.hkust-gz.edu.cn; wchen361@connect.hkust-gz.edu.cn; hlei573@connect.hkust-gz.edu.cn; eewanglj@connect.hkust-gz.edu.cn; haoang.li.cuhk@gmail.com). (Haoxuan Xu and Tianfu Li contributed equally to this work.) (Corresponding author: Haoang Li.)

Yi Liu is with Tsinghua University, Shenzhen 518055, China (e-mail: yiliu24@mails.tsinghua.edu.cn).

Jin Wu is with University of Science and Technology Beijing, Beijing 100083, China (e-mail: wujin@ustb.edu.cn).

Yunfan Lou is with National University of Singapore, 119077, Singapore (e-mail: e1373933@u.nus.edu).

Hesheng Wang is with Shanghai Jiao Tong University, Shanghai 200240, China (e-mail: wanghesheng@sjtu.edu.cn).

stop in front of them". A conventional agent may bypass the target group entirely because it cannot semantically recognize their ongoing activity and thus misclassifies them as obstacles (Fig. 1(b)). Conversely, even if it approaches the group, it may fail to anticipate their personal space and future motion, resulting in an intrusive or even colliding trajectory (Fig. 1(c)). These failures reveal a fundamental gap in current VLN: the lack of explicit human-centric reasoning, which requires both semantic understanding of human activities and geometric foresight of human motion.

To address this gap, we propose **HCSG**, a **H**uman-**C**entric **S**emantic-**G**eometric Reasoning framework for VLN that explicitly equips the agent with semantic understanding and geometric foresight. The central idea is that successful navigation in human-populated environments requires not only recognizing where people are, but also reasoning about what they are doing, how they may move, and how the robot should behave with respect to them under the language goal. Accordingly, HCSG departs from conventional VLN agents that react to single-frame observations or rank navigable candidates primarily from static environmental cues [1], [2], [18]. Instead, once humans are detected, the agent pauses briefly to collect a short temporal observation window, providing dynamic context for subsequent human-aware reasoning.

Based on this observation sequence, HCSG processes human information through two complementary streams. To address the semantic deficit illustrated in Fig. 1(b), the **Semantic Stream** leverages a large vision-language model (VLM) to analyze cropped human-centered observations and produce explicit descriptions of actions, intentions, and social context. These semantic representations enable the agent to move beyond treating people as undifferentiated obstacles, and instead align observed human behavior with the instruction, for example by identifying *a group of people discussing* as the intended interaction target. To address the geometric deficit illustrated in Fig. 1 (c), the **Geometric Stream** models human dynamics through fine-grained pose cues and future trajectory forecasting. By encoding short-term temporal observations, this stream captures spatio-temporal patterns of motion and anticipates future human occupancy, providing predictive signals for proactive collision avoidance. Moreover, pose structure offers complementary physical evidence for behavior understanding, such as body orientation and coordinated group activity, thereby grounding the semantic interpretation in observable motion patterns.

The resulting semantic and geometric cues are fused into the agent's topological representation, so that human-aware states become explicitly available to the navigation policy rather than remaining buried in global scene features. This enriched representation allows a cross-modal transformer to align the global linguistic goal with localized human-centric nodes and select waypoints that are not only goal-directed, but also socially appropriate. To further enforce safe interaction, we introduce a **Social Distance Loss** that penalizes trajectories violating predicted human occupancy and interpersonal space. This objective combines strict collision avoidance with softer repulsion from future human regions, encouraging the agent to respect socially comfortable distances during approach and navigation. In this way, HCSG moves beyond passive obstacle avoidance toward instruction-conditioned human-centric navigation in dynamic environments.

Extensive experiments on the HA-VLNCE benchmark [10] demonstrate that HCSG achieves state-of-the-art performance, improving Success Rate by 14.3% and reducing Collision Rate by 34.5% on the challenging validation-unseen split.

In summary, the contributions of this work are four-fold:

- We formulate human-centric VLN as a navigation problem that requires both semantic understanding of human activities and geometric foresight of human motion, moving beyond the conventional view of pedestrians as mere obstacles.
- We propose HCSG, a dual-stream human-aware reasoning framework that combines VLM-based semantic interpretation with pose- and trajectory-based motion modeling for instruction-conditioned navigation in dynamic social environments.
- We introduce a Social Distance Loss to discourage trajectories that violate predicted human occupancy and interpersonal space, promoting safe and socially compliant navigation.
- HCSG achieves state-of-the-art performance on the HA-VLNCE benchmark, especially in success rate and collision rate.

II. Related Work

A. Vision-Language Navigation

VLN requires an embodied agent to interpret natural language instructions together with egocentric visual observations in order to reach a goal in either discrete [3] or continuous environments [5]. Early work relied on sequence-to-sequence architectures, including recurrent neural network (RNN) [19]–[21], Long Short Term Memory network (LSTM) [3], and Transformer-based models [22]. Subsequent advances improved VLN along several axes: more effective learning strategies [23], [24], data-augmentation techniques [25]–[29], large-scale pre-training [30]–[34], LLM-based methods [6], [8] instruction-aware semantic enhancement [35] and language-driven spatial localization [36]. To strengthen spatial reasoning, DUET [37] and ETPNav [2] employed topological maps that record the agent's trajectory for global planning, while BEVBert [1] and GridMM [38] generated bird's-eye-view (BEV) grid maps for geometry-consistent scene understanding; VER [39] further extended this idea to 3D voxel grids. However, these methods largely focus on scene understanding in static environments and provide limited support for reasoning about dynamic human activity, especially in indoor spaces where pedestrians may appear, move, and interact unpredictably.

B. Human-aware Understanding

Once navigation moves beyond static scenes, a key challenge is no longer only where to go, but also how to interpret the humans encountered along the way. Human-aware understanding therefore ranges from fine-grained human attribute

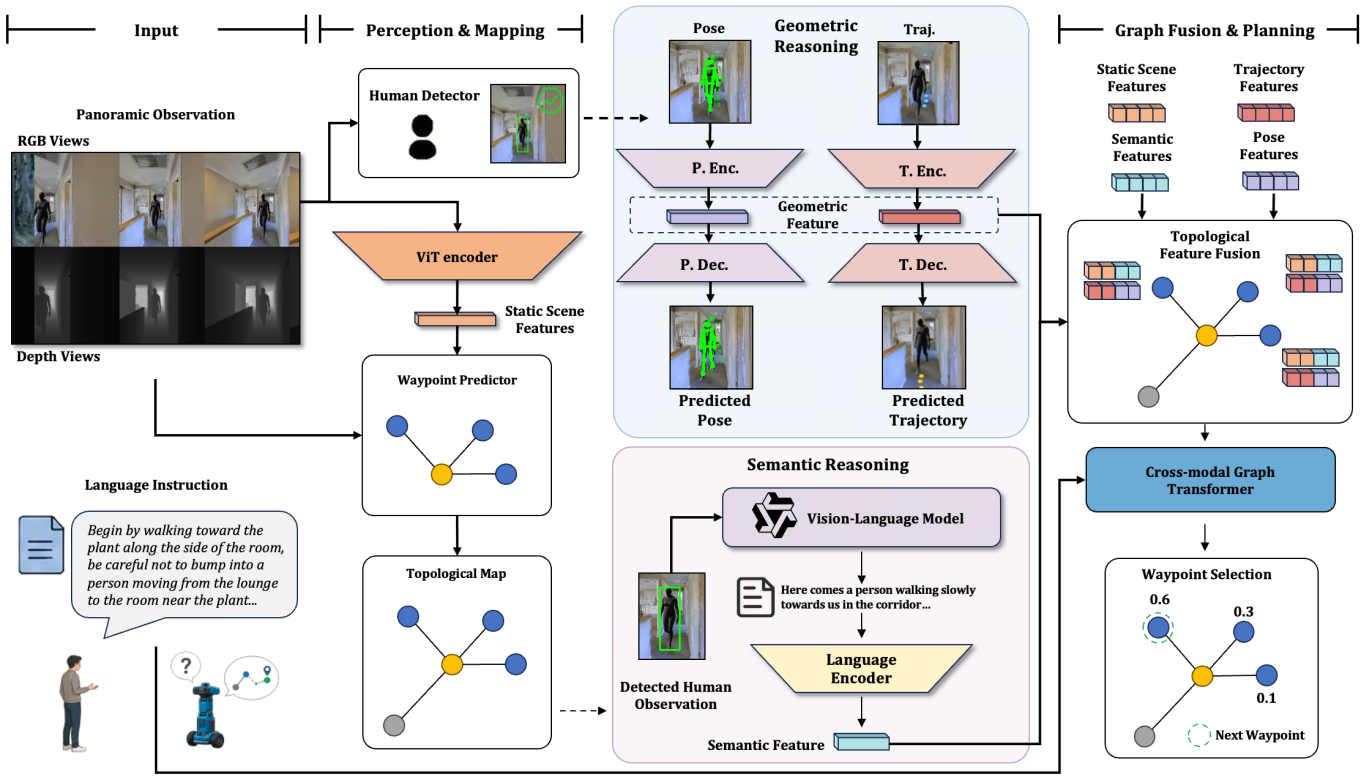


Fig. 2. Overview of the proposed Human-Centric Semantic-Geometric Reasoning framework for Vision-Language Navigation (HCSG). Starting from panoramic observations and a language instruction, the agent performs perception and mapping, followed by parallel **Geometric Reasoning** and **Semantic Reasoning** for detected humans. The resulting human-centric features are fused with static scene features in the topological graph, which is then processed by a Cross-modal Graph Transformer for language-guided planning, i.e. waypoint selection.

grounding [40] to modeling human motion, actions, and intentions. Trajectory-centric pedestrian models predict motion and intent [41], key-point pose predictors encode skeletal dynamics [42], and large-scale foundation models further distill the semantics of continuous human motion [43], together supporting reasoning about dynamic human-environment interactions. Yet within the VLN literature, current policies still make limited use of human-aware cues for modeling actions and intentions. Projects such as DRAGON [14] and CoRI [15] let robots infer user goals indirectly through dialogue, and Social-LLaVA [16] and VLM-Social-Nav [17] deploy vision-language models to detect pedestrian intent for collision avoidance; however, these efforts remain focused on simply steering clear of people. Genuine human awareness entered VLN research only with the release of the HA-VLN [44] and HA-VLNCE [10] datasets, but their baseline agents still view pedestrians as moving obstacles and lack deeper insight into human actions and intentions. Therefore, the next step for VLN is to move beyond treating humans as mere moving obstacles, toward explicitly understanding human actions and intentions so that agents can ground instructions more accurately in dynamic human-populated environments.

C. Social Robot Navigation

Beyond understanding human behavior, robots must also respond to it in socially appropriate ways. Social robot navigation therefore extends obstacle avoidance by requiring agents to follow implicit social norms, such as maintaining comfortable interpersonal distances and producing legible motion. From a control perspective, geometric and optimization-based approaches model social compliance through hybrid navigation architectures, human-aware costs, elastic-band refinement, and short-horizon pedestrian motion prediction [45]–[49]. Other studies further emphasize legibility and predictability as key properties of socially appropriate motion in multi-agent environments [50]. From a learning perspective, Deep Reinforcement Learning and risk-aware Model Predictive Control have been used to model complex multi-agent interactions under predictive uncertainty [51]–[53]. Despite these advances, most social navigation systems remain task-agnostic, treating pedestrians as generic dynamic obstacles to be avoided or politely bypassed. Yet social compliance alone is insufficient for VLN. While human-aware understanding provides cues about human actions and intentions, and social navigation provides principles for socially appropriate response, neither alone supports instruction-conditioned human-centric navigation. An agent must understand who a person is in relation to the language goal, what that person is doing, and how to navigate around or toward them appropriately. This gap motivates our HCSG framework.

III. Problem Formulation

Our work mainly follows the setup of Vision-Language Navigation in Continuous Environments (VLN-CE). For most

VLN-CE works [1], [2], [38], the task requires an embodied agent $\mathcal{A}$ to navigate in 3D environments $\mathcal{E}$, guided by natural language instruction $\mathcal{I}$. Formally, at each timestep $t \in [0, T]$: $\mathcal{A}$ receives spherical observation $\mathbb{O}_t = (\mathcal{V}_t^{\text{rgb}}, \mathcal{V}_t^{\text{depth}})$ with $\mathcal{V}_t^{\text{rgb}} = \{v_{t,k}^{\text{rgb}}\}_{k=1}^{12}$, $\mathcal{V}_t^{\text{depth}} = \{v_{t,k}^{\text{depth}}\}_{k=1}^{12}$ covering 360° FoV at 30° intervals. The agent learns a policy $\pi : (\mathbb{O}_t, \mathcal{I}) \rightarrow \mathbf{a}_{t+1}$ which outputs $\mathbf{a}_{t+1} \in \mathbb{R}^3$ representing relative displacement $(\Delta x, \Delta y, \Delta\theta)$ in SE(2) space. The navigation trajectory $\tau = \{\mathbf{p}_t\}_{t=0}^{T}$ terminates when $\|\mathbf{p}_T - \mathbf{p}_{\text{goal}}\|_2 < \delta_{\text{th}}$. Additionally, in our human-aware VLN-CE task setting, the observations $\mathbb{O}_t$ of the agent include dynamic humans, and the instructions $\mathcal{I}$ also contain descriptions of people in the scene.

IV. Methodology

Based on the standard waypoint-based navigation policy [1], [2], [18], we augment the conventional VLN framework with a human-centric reasoning pipeline. As illustrated in Fig. 2, the overall navigation process first builds an online topological graph from panoramic observations and candidate waypoints, and then injects human-centric features into the graph whenever humans are detected. We describe this integration mechanism in Sec. IV-A.

Once a human is detected at a waypoint, the agent activates two complementary reasoning streams. The Geometric Reasoning Module, detailed in Sec. IV-B, estimates observed human pose and trajectory and further forecasts their future evolution. The Semantic Reasoning Module, detailed in Sec. IV-C, leverages a VLM to infer human actions, intentions, and navigation-relevant social context. The resulting semantic-geometric features are fused into the online topological representation for language-guided waypoint selection. Finally, the Social Distance Loss in Sec. IV-D regularizes the navigation policy toward collision-free and socially compliant behavior.

A. Navigation with Human Reasoning

In this subsection, we describe how human-centric reasoning is integrated into the standard waypoint-based VLN pipeline. As illustrated in Fig. 2, our method injects human features into the topological graph during navigation upon human detection. We next detail the concrete mechanism for integrating human reasoning into the navigation process.

Waypoint-based Navigation. At timestep t, the agent receives a panoramic observation $\mathbb{O}_t = (\mathcal{V}_t^{\text{rgb}}, \mathcal{V}_t^{\text{depth}})$ and passes it to an external waypoint predictor f_{way} [2] to estimate navigable positions, generating candidate nodes $\mathcal{W}_t$ for the next action. Meanwhile, the pre-trained visual encoder extracts features from each perspective and fuses them into the corresponding nodes estimated by the waypoint predictor:

$$\mathcal{F}_{t,k}^{\text{static}} = \mathcal{E}_v\left(\mathbb{O}_{t,k}\right), \tag{1}$$

where k denotes the panoramic perspective index, $\mathcal{E}_v$ denotes the pre-trained visual encoder. These features are treated as static, since they are all captured at a single instant. Building upon this, our approach explicitly incorporates the dynamic features of nearby people. The agent then reasons over the structure formed by the waypoint predictor and the representation from the feature encoder, and commits to one node as its next action.

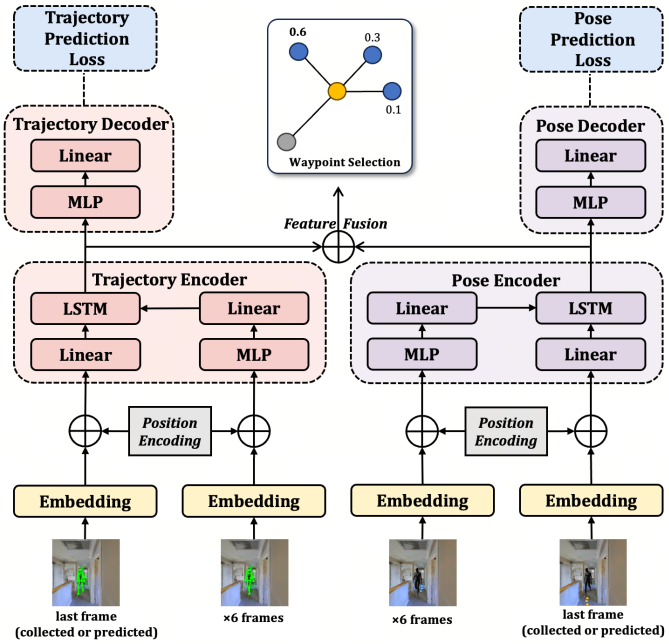


Fig. 3. Pipeline of the Human Geometric Reasoning Module. We decompose the geometric motion of the human body into two complementary components: trajectory and pose. Both the trajectory and pose predictors utilize an LSTM-based encoder-decoder structure to process the input frame embeddings. By predicting these components and fusing intermediate features extracted from the network, we obtain future-oriented geometric representations. These fused features are subsequently supplied to the downstream navigation module for tasks such as waypoint selection.

Human Detection. When the agent reaches a new waypoint, it needs to wait for the waypoint predictor to estimate new navigable areas. We leverage this property of waypoint-based navigation to design a short temporal observation strategy. During this brief pause, a human detector $\mathcal{D}$ analyzes the panorama $\mathcal{V}_t^{\text{rgb}}$ and returns a set of bounding boxes $\mathcal{H}_t^{\text{det}}$ for humans. If no human is detected, the agent will skip the remaining steps and proceed normally, selecting its next waypoint from $\mathcal{W}_t$ according to the navigation policy $w_{t+1} = \pi(\mathbb{O}_{<=t}, \mathcal{I}, \mathcal{H}_{<=t})$, where $\mathcal{H}_{<=t}$ represents the features collected related to humans up to this point. When humans are detected, the agent continues to use this brief pause to acquire an observational sequence $\mathcal{S} = \langle \mathbb{O}_\tau \rangle_{\tau=t}^{t+m-1}$ to obtain dynamic information, where m denotes the gathered timesteps.

Human Reasoning. Upon detecting nearby humans, the agent reorganizes the temporal observation sequence $\mathcal{S}$ into individual-centric trajectories $\mathcal{P}_j^{1:m}$. To achieve holistic human understanding, we process each trajectory through two specialized streams. The geometric encoder captures spatio-temporal motion patterns for future occupancy reasoning, while the semantic encoder leverages a VLM to infer social context and human intent. Formally,

$$\mathcal{F}_{t,k,j}^{\text{geo}} = \mathcal{E}_g\left(\mathcal{P}_j^{1:m}\right), \tag{2}$$

$$\mathcal{F}_{t,k,j}^{\text{sem}} = \mathcal{E}_s\left(\mathcal{P}_j^{1:m}\right), \tag{3}$$

where $\mathcal{E}_g$ and $\mathcal{E}_s$ denote the geometric and semantic encoders, t denotes the navigation timestep, k denotes the panoramic

perspective index, and j denotes the detected human index. Detailed architectures are provided in Sec. IV-B and Sec. IV-C.

Topological Feature Fusion. To enable the navigation policy to reason over human dynamics, we inject the synchronized human features into the agent's topological representation. Specifically, we use a fusion layer to aggregate the static scene features $\mathcal{F}_{t,k}^{\text{static}}$ with the mean dynamic representations of all J humans associated with a specific waypoint node:

$$\mathcal{F}_{t,k}^{\text{fused}} = \text{MLP}\left(\mathcal{F}_{t,k}^{\text{static}}, \frac{1}{J}\sum_{j=1}^{J}(\mathcal{F}_{t,k,j}^{\text{geo}} + \mathcal{F}_{t,k,j}^{\text{sem}})\right), \quad (4)$$

This human-aware topological node feature $\mathcal{F}_{t,k}^{\text{fused}}$ allows the Cross-modal Graph Transformer to align the global linguistic instruction $\mathcal{I}$ with localized, human-centric constraints. The final action a_{t+1} is selected by evaluating candidate waypoints through a Feed-Forward Network (FFN), ensuring the trajectory is both goal-directed and socially appropriate.

$$a_{t+1} = \text{FFN}(\text{GASA}\left(\mathcal{F}_t^{\text{fused}}, \mathcal{W}_t, \mathcal{I}\right)). \quad (5)$$

Having described how human-centric reasoning is integrated into the navigation process, we next detail the construction of the geometric and semantic human-centric features in Sec. IV-B and Sec. IV-C, followed by the social distance loss in Sec. IV-D.

B. Geometric Reasoning

Following the integrated reasoning pipeline described in Sec. IV-A, the Geometric Reasoning module specifically addresses the geometric deficit by forecasting fine-grained pose and coarse-grained trajectory. For human-centric geometric reasoning, we decompose human motion into two complementary cues: body posture and movement trajectory, as shown in Fig. 3. Body posture provides fine-grained cues about the activity in which a person is currently engaged, while movement trajectory offers coarse-grained evidence about where the person may move next. Consequently, our module is designed to interpret human behavior precisely through these two facets via estimation and prediction.

Observed Geometric State Estimation. For the collected sequence $\mathcal{S}$, when we organize the information by human index j, we extract the keypoints data of the human body to estimate the pose:

$$\mathcal{P}_j^{\text{pose}} = \left\{\mathbf{K}_j^m\right\}_{m=1}^{M}, \quad \mathbf{K}_j^m = \left\{\mathbf{k}_{i,m}^{\text{pose}}\right\}_{i=1}^{17} \in \mathbb{R}^{17\times 3}, \quad (6)$$

where $\mathbf{k}_{i,m}^{\text{pose}} = (x_{i,m}^{\text{kp}}, y_{i,m}^{\text{kp}}, c_{i,m}^{\text{conf}})$ denotes the 2D joint position and detection confidence of the i-th keypoint at frame m, and M denotes the observation length.

Simultaneously, we estimate the relative human position $\mathcal{P}_j^{\text{traj}}$ by back-projecting the depth value at the human bounding-box center into 3D space using the standard pinhole camera model and the camera intrinsic matrix [54]. The resulting 3D point is then transformed from the camera frame to the agent-centric frame according to the current agent heading, and its horizontal coordinates are taken as the trajectory cue for downstream geometric reasoning.

Future Geometric State Forecasting. Once the data are reorganized, the agent predicts the future pose and trajectory of each person. As shown in Fig. 3, both the pose and the trajectory predictors share an underlying architecture composed of two core components. An LSTM-based encoder distills salient motion patterns from sequential inputs. Furthermore, a lightweight decoder propagates state predictions frame-by-frame. The features output by the encoder module will be fused and embedded to the graph representation for subsequent navigation module.

Crucially, the pose and trajectory prediction losses are computed on every step and their gradients back-propagated to fine-tune the model while the agent navigates on the train set. Concretely, the pose prediction loss combines coordinate and confidence accuracy:

$$\mathcal{L}_{\text{pose}} = \frac{1}{N}\sum_{i=1}^{N}\left(\|\mathbf{k}_i^{\text{pred}} - \mathbf{k}_i^{\text{gt}}\|_2^2 + \gamma_1(c_i^{\text{pred}} - c_i^{\text{gt}})^2\right), \quad (7)$$

with γ_1 regulating confidence importance. And trajectory prediction integrates position and velocity constraints:

$$\mathcal{L}_{\text{traj}} = \frac{1}{T}\sum_{t=1}^{T}\left(\|\mathbf{t}_t^{\text{pred}} - \mathbf{t}_t^{\text{gt}}\|_2^2 + \gamma_2\|\mathbf{v}_t^{\text{pred}} - \mathbf{v}_t^{\text{gt}}\|_2^2\right), \quad (8)$$

as training progresses, the relative weights of these two losses in the total objective are progressively annealed.

Future-oriented Geometric Feature Construction. The forecasting task is designed not only to predict future human states, but also to learn intermediate representations that contain future-aware motion information. We therefore use the hidden representations from the trajectory and pose forecasting branches as the geometric feature $\mathcal{F}_{t,k,j}^{\text{geo}}$ for human j. Since these representations are learned under explicit future prediction supervision, they are encouraged to capture imminent motion tendencies and potential future occupancy. The resulting geometric feature is then passed to the topological feature fusion module in Sec. IV-A, where it is combined with semantic human features and static scene features for downstream waypoint selection. The effectiveness of this future-oriented design is evaluated in the ablation studies.

C. Semantic Reasoning

While geometric forecasting ensures the basic understanding and physical safety, it lacks the capability to interpret the high-level semantics of human activities, which is essential for grounding instructions such as *"wait for the person calling"*. To bridge this gap, we design a Semantic Reasoning stream that leverages the zero-shot reasoning capability of Large Vision-Language Models.

Visual-to-Linguistic Interpretation. For each detected human j at timestep t, we first extract and crop the corresponding perspective view from the panoramic observation V_t^{rgb} based on the human detector. This image $\mathcal{I}_{t,k,j}$ serves as the visual input to the semantic branch. As illustrated in Fig. 4, the goal of this branch is to infer human activity, intent, and navigation-relevant social context, thereby complementing the geometric cues with higher-level semantics. To guide the

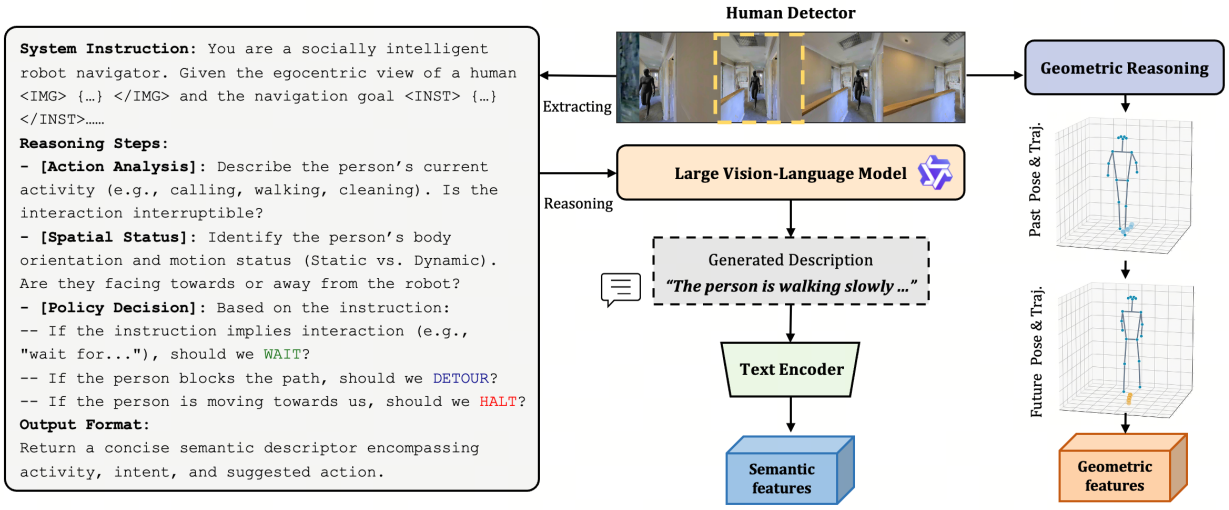


Fig. 4. Illustration of human-centric semantic-geometric reasoning. Given the human detected in the image sequence, the agent performs semantic reasoning to infer activity, intent, and navigation-relevant social context, while geometric reasoning captures trajectory and pose cues, including their temporal evolution. These complementary cues provide structured human-aware representations for downstream navigation.

Vision-Language Model toward navigation-relevant reasoning rather than generic captioning, we formulate a task-specific prompt $\mathcal{Q}$. Conditioned on $\mathcal{I}_{t,k,j}$ and $\mathcal{Q}$, the VLM generates a natural-language description $\mathcal{T}_{t,k,j}$:

$$\mathcal{T}_{t,k,j} = \text{VLM}(I_{t,k,j}, \mathcal{Q}). \tag{9}$$

For instance, the model might output *"A person is standing still and sorting clothes"* or *"A person is walking rapidly towards the corridor"*. This linguistic representation explicitly captures the high-level semantic intent that the geometric reasoning module fails to convey.

Semantic Feature Encoding. To integrate this linguistic insight into the navigation policy, we instantiate the semantic encoder $\mathcal{E}_s$ with a pre-trained text encoder (i.e. CLIP [55]), which maps the generated description $\mathcal{T}_{t,k,j}$ into a high-dimensional semantic representation:

$$\mathcal{F}^{\text{sem}}_{t,k,j} = \mathcal{E}_s(\mathcal{T}_{t,k,j})\,. \tag{10}$$

Here, $\mathcal{E}_s$ denotes the text encoding stage of the semantic branch. Unlike the geometric feature $\mathcal{F}^{\text{geo}}_{t,k,j}$, which encodes *how* a person moves, $\mathcal{F}^{\text{sem}}_{t,k,j}$ captures *why* the person acts, providing crucial context for instruction alignment. Finally, this semantic representation is fused with the geometric features, enabling the Cross-modal Graph Transformer to reason jointly over physical constraints and social context.

D. Social Distance Losses

Beyond the reasoning modules, we design a unified Social Distance Loss that discourages the planner from outputting any trajectory that would intrude upon space already occupied by humans, thereby ensuring collision-free motion and safe human-robot interaction. It is enforced through two complementary mechanisms:

Actual Collision Loss. During navigation, each collision is penalized by:

$$\mathcal{L}_{\text{coll}} = \lambda_c \sum_{i \in \mathcal{C}} \delta_i, \tag{11}$$

where $\mathcal{C}$ is the set of collision events and δ_i is the penalty coefficient (empirically set to 3.0).

Proximity Avoidance Loss. For humans detected within safety radius $r_s = 1.0\,\text{m}$, we compute a repulsive loss that grows as the robot approaches the human:

$$\mathcal{L}_{\text{prox}} = \lambda_p \sum_j \frac{\phi(\theta_j)}{\max(\|\mathbf{d}_j\|_2^2, \epsilon_p)}, \tag{12}$$

where $\mathbf{d}_j = (d_x^j, d_y^j)$ is the relative position vector, $\phi(\cdot)$ is the front-facing penalty weighting function ($[0.25, 1.0]$), and $\epsilon_p = 0.0625\,\text{m}^2$ prevents divergence.

Finally, combining the auxiliary forecasting losses in Eqs. (7) and (8), the social safety losses in Eqs. (11) and (12), and the standard navigation imitation objective, the overall training objective is formulated as:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{pose}} + \mathcal{L}_{\text{traj}} + \mathcal{L}_{\text{coll}} + \mathcal{L}_{\text{prox}} + \mathcal{L}_{\text{nav}}, \tag{13}$$

where $\mathcal{L}_{\text{nav}} = -\mathbb{E}[\log P(a^*|s)]$ denotes the standard cross-entropy loss for navigation action prediction.

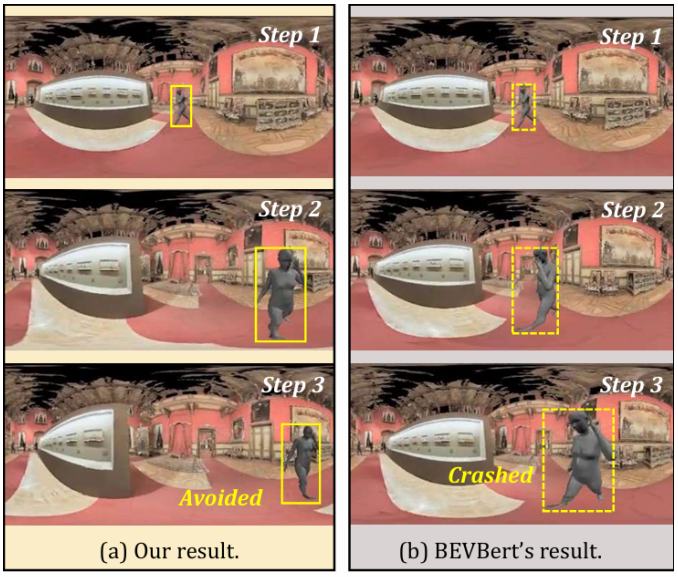
V. Experiments

A. Experiment Setup

1) Dataset: The HA-VLNCE dataset integrates 486 SMPL-format 3D human motion models spanning 172 daily activity categories (including running, climbing, and phone conversations, totaling 58,320 frame sequences), annotates 910 dynamic human instances across 90 scenes (with up to 10 humans per scene), and extends 16,844 human-centric instructions averaging 15 words in length, comprising a training set of 10,819 instructions for learning human dynamic interactions and path planning, along with validation sets containing 778 seen validation set for evaluating generalization in familiar environments and 1,839 unseen validation set testing adaptability in novel scenarios, collectively establishing a continuous human-aware VLN benchmark for assessing agent performance in human-populated environments [10].

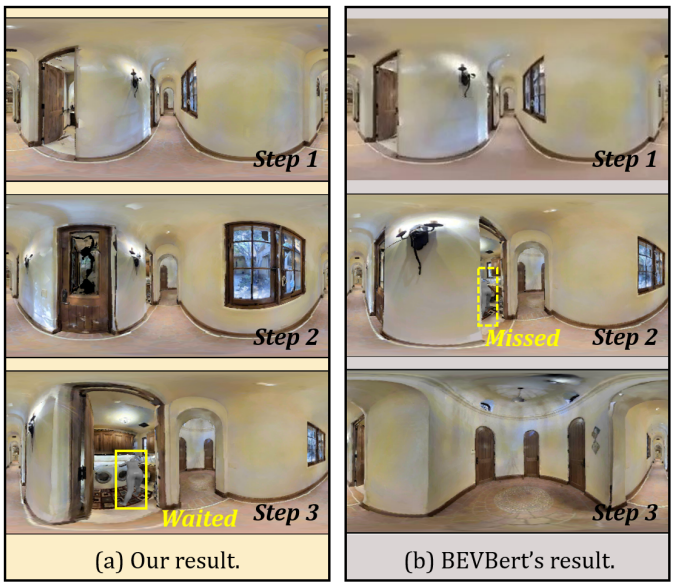
TABLE I
COMPARISON WITH STATE-OF-THE-ART METHODS ON THE HA-VLNCE BENCHMARK. BEST RESULTS ARE HIGHLIGHTED IN BOLD AND SECOND-BEST RESULTS ARE UNDERLINED. FOR CR AND SR, WE ADDITIONALLY REPORT THE RELATIVE IMPROVEMENT OF OUR METHOD OVER THE SECOND-BEST BASELINE.

Models	Validation Seen			Validation Unseen		
	TCR↓	CR↓	SR↑	TCR↓	CR↓	SR↑
HA-VLN-CMA [10]	63.09	0.77	0.05	47.06	0.77	0.07
HA-VLN-CMA-DA [10]	17.45	0.61	0.17	27.25	0.69	0.09
HA-VLN-VL [10]	4.44	0.52	0.20	6.63	0.59	0.14
LAW-VLNCE [56]	4.31	0.54	0.21	5.88	0.65	0.15
DUET [37]	4.18	0.48	0.22	5.74	0.63	0.16
ETPNav [2]	4.07	<u>0.43</u>	0.24	6.94	0.58	0.17
GridMM [38]	3.92	0.45	0.24	5.76	0.59	0.18
BEVBert [1]	<u>3.64</u>	0.46	<u>0.27</u>	**4.71**	<u>0.55</u>	<u>0.21</u>
HCSG (Ours)	**3.63**	**0.34** (↓20.9%)	**0.29** (↑7.4%)	<u>5.02</u>	**0.36** (↓34.5%)	**0.24** (↑14.3%)



Instruction: *Begin by ..., and continue moving through the lounge, where the person on the phone continues their conversation while pacing ... stop at the entrance to ...*

Fig. 5. Visualization of results on HA-VLNCE benchmark. (a) shows that our model avoided the person and continued to follow the instruction, while the SOTA model (BEVBert) in (b) collided with the person.



Instruction: *Begin your journey ... This individual might be sorting clothes, operating the washing machine, or folding freshly loundered items. Your task is to wait ...*

Fig. 6. Visualization of results on HA-VLNCE benchmark. (a) shows that our model recognized the human behavior in the language description and successfully completed the VLN task, while the SOTA model (BEVBert) in (b) missed this person.

2) Evaluation Metrics: Following HA-VLNCE [10], we evaluate the agent using navigation and human-safety metrics, including Navigation Error (NE), Success Rate (SR), Total Collision Rate (TCR), and Collision Rate (CR). NE measures the average distance between the agent's final position and the target, while SR measures the proportion of episodes successfully completed without collision. TCR reflects the overall frequency of human-related collisions throughout navigation, and CR measures the proportion of episodes that contain at least one human-related collision. Since this work focuses on human-populated navigation, our main comparison emphasizes SR, TCR, and CR, which directly reflect task completion and social safety in dynamic human environments. NE is included in the evaluation protocol for completeness, but the main tables prioritize safety-related metrics.

3) Implementation Details: All experiments are conducted using PyTorch framework on two NVIDIA A6000 GPUs. We pre-train the model in static environments [5] for 100,000 iterations with a batch size of 64 and a learning rate of 5×10^{-5}, employing the AdamW optimizer. During fine-tuning, the agent interacts with the environments online via the Habitat Simulator [12]. At each waypoint, we apply YOLO-Pose [57] to perform human detection; once a person is detected, the agent collects 6 frames of information and predicts the trajectories and poses for the next 3 frames. We employ Qwen3-VL-2B-Instruct [58] to reason human motion. Fine-tuning is performed for 15,000 iterations with a batch

TABLE II
ABLATION STUDY ON GEOMETRIC REASONING.

Pose	Trajectory	Validation Seen			Validation Unseen		
		TCR↓	CR↓	SR↑	TCR↓	CR↓	SR↑
✗	✗	3.98	0.39	0.26	6.09	0.48	0.20
✓	✗	3.91	0.38	**0.29**	5.92	0.47	0.22
✗	✓	**3.72**	**0.36**	0.27	5.31	0.41	0.22
✓	✓	**3.72**	**0.36**	0.28	**5.21**	**0.40**	**0.23**

TABLE III
ABLATION STUDY ON VLM-BASED SEMANTIC REASONING.

Setting	Validation Seen			Validation Unseen		
	TCR↓	CR↓	SR↑	TCR↓	CR↓	SR↑
w/o VLM	3.72	0.36	0.28	5.21	0.40	0.23
w VLM	**3.66**	**0.34**	**0.29**	**5.02**	**0.36**	**0.24**

size of 8 and a learning rate of 1×10^{-5}.

B. Comparisons with State-of-the-art

As evidenced in Table I, HCSG establishes a new state-of-the-art on the HA-VLNCE dataset. A critical observation from the baseline methods is the inherent trade-off between navigation success and safety; traditional agents tend to improve goal-reaching performance without explicitly controlling human-related collisions. However, our method effectively overcomes this limitation. Under the challenging validation-unseen split, HCSG reduces the critical Collision Rate (CR) to 0.36, yielding a significant reduction of 34.5% compared to BEVBert and 37.9% compared to ETPNav. Concurrently, our agent attains a Success Rate (SR) of 0.24, surpassing BEVBert (0.21) and ETPNav (0.17). This simultaneous improvement in both success and safety metrics demonstrates that explicitly modeling human behaviors allows the agent to navigate aggressively yet safely, rather than simply stopping or colliding with pedestrians when encountering pedestrians.

Fig. 5 and Fig. 6 provide qualitative visualizations of our model's performance on the HA-VLNCE benchmark. In Fig. 6, the instruction requires the agent to wait for a person sorting clothes. Traditional methods fail to capture this semantic cue and attempt to bypass the person immediately, leading to a collision or task failure. In contrast, our HCSG correctly interprets the activity description, halts at a safe distance, and resumes navigation once the path is clear. Similarly, Fig. 5 demonstrates a dynamic avoidance scenario where the agent anticipates the trajectory of a pacing person. These visualizations qualitatively validate that our framework succeeds not only by reacting to obstacles but by understanding the semantic intent and future geometry of human motion.

C. Ablation Studies

1) Significance of Geometric Reasoning: We first analyze the impact of the explicit geometric reasoning components, as detailed in Table II. The baseline model without any specific human understanding modules yields a high collision rate of 0.48 in unseen environments. Incorporating trajectory features alone lowers the collision rate significantly to 0.41. This result indicates that predicting future occupancy helps the agent proactively plan paths that avoid dynamic obstacles. On the other hand, adding pose features alone results in a noticeable improvement in Success Rate, increasing from 0.20 to 0.22. This suggests that pose information serves as a fine-grained semantic signal, helping the agent distinguish between different human activities relevant to the instruction. The synergistic fusion of both pose and trajectory features achieves the optimal performance with a CR of 0.40 and SR of 0.23. This confirms that geometric constraints and behavioral semantics are complementary; the former ensures physical safety while the latter aids in task grounding.

TABLE IV
ABLATION STUDY ON FUTURE VS. PAST-ORIENTED FEATURE.

Feature	Validation Seen			Validation Unseen		
	TCR↓	CR↓	SR↑	TCR↓	CR↓	SR↑
Past-oriented	3.81	0.37	0.27	5.25	0.41	0.21
Future-oriented	**3.72**	**0.36**	**0.28**	**5.21**	**0.40**	**0.23**

TABLE V
ABLATION STUDY ON FINE-TUNING VS. FIXED-PARAMETER DESIGN.

Setting	Validation Seen			Validation Unseen		
	TCR↓	CR↓	SR↑	TCR↓	CR↓	SR↑
Fixed-parameter	3.86	0.39	0.27	5.27	0.42	0.22
Fine-tuning	**3.72**	**0.36**	**0.28**	**5.21**	**0.40**	**0.23**

2) Impact of VLM-based Semantic Reasoning: To further verify the necessity of the Semantic Stream, we conduct an ablation study on the VLM component, as shown in Table III. The model without VLM support relies solely on geometric cues for human avoidance. While it achieves reasonable safety, the integration of VLM-based descriptions further optimizes performance. Specifically, the inclusion of VLM reduces the Collision Rate from 0.40 to 0.36 and improves the Success Rate from 0.23 to 0.24 in unseen environments. This improvement can be attributed to the VLM's capability to interpret complex social scenarios that geometric features cannot capture, such as identifying whether a person is stationary and interacting with an object or about to turn around. By converting visual human actions into linguistic descriptions, the VLM aligns the visual observations more effectively with the natural language instructions, leading to safer and more intelligent decision-making.

3) Future vs. Past-oriented Feature: In the previous section, we emphasize that we employ a future-oriented understanding approach, which involves having our agent predict the future state of humans to understand its current ongoing activities. We argue that this method is better than the past-oriented understanding approach of directly encoding the collected information, because only by knowing what a person will do in the future can it be said that one truly understands his behavior. Empirical results in Table IV support this hypothesis. The future-oriented design outperforms the past-oriented counterpart in unseen scenarios, specifically increasing the Success

TABLE VI
ABLATION ON SOCIAL DISTANCE LOSS.

$\mathcal{L}_{coll}$	$\mathcal{L}_{prox}$	Validation Seen			Validation Unseen		
		TCR↓	CR↓	SR↑	TCR↓	CR↓	SR↑
✗	✗	4.06	0.42	0.26	6.73	0.55	0.20
✓	✗	4.02	0.41	0.26	6.35	0.52	0.21
✗	✓	3.99	0.39	0.27	6.13	0.50	0.21
✓	✓	**3.72**	**0.36**	**0.28**	**5.21**	**0.40**	**0.23**

TABLE VII
ABLATION ON RGB/DEPTH INPUTS.

RGB	Depth	Validation Seen			Validation Unseen		
		TCR↓	CR↓	SR↑	TCR↓	CR↓	SR↑
✓	✗	5.93	0.51	0.20	7.31	0.56	0.15
✗	✓	4.48	0.40	0.26	7.02	0.46	0.21
✓	✓	**3.72**	**0.36**	**0.28**	**5.21**	**0.40**	**0.23**

Rate from 0.21 to 0.23 and decreasing the Collision Rate from 0.41 to 0.40. This confirms that predictive features provide more actionable guidance for navigation than historical encodings.

4) Fine-tuning vs. Fixed-parameter Design: We further claim that the continuous fine-tuning strategy for the geometric reasoning module provides more adaptive human features that are more adaptable than fixed pre-trained features. To validate this, we compared our adaptive approach against a fixed-parameter baseline in Table V. The quantitative results confirm the advantage of adaptation. Specifically, the fine-tuning mechanism reduces the Collision Rate from 0.42 to 0.40 and improves the Success Rate from 0.22 to 0.23 in validation unseen environments. This improvement implies that the domain distribution of human behaviors in the navigation environment differs from the pre-training dataset. Consequently, training-time adaptation allows the motion forecasting module to adapt to specific movement patterns encountered during the episode, thereby providing more precise features for the navigation policy.

5) Significance of Social Distance Loss: As demonstrated in Table VI, the proposed dual loss mechanism exhibits complementary advantages in optimizing navigation safety. The collision penalty loss significantly reduces the Collision Rate from 0.55 to 0.52 in unseen environments. Concurrently, the proximity avoidance loss achieves a reduction in Total Collision Rate from 6.73 to 6.13. Crucially, their combined implementation creates a synergistic prevention-penalty mechanism that drives substantial further improvements. The combination lowers the Unseen Collision Rate to 0.40 and optimizes the Total Collision Rate to 5.21, validating the necessity of enforcing both immediate collision constraints and proactive spacing.

6) Ablation on RGB/Depth inputs: Table VII confirms the robustness of our model to different visual modalities. The model relying solely on RGB inputs suffers from a high Collision Rate of 0.56 and a low Success Rate of 0.15 in unseen scenes, as it lacks precise depth information to estimate distances to pedestrians. Integrating Depth information significantly alleviates this issue, lowering the Collision Rate to 0.46 and boosting the Success Rate to 0.21. However, the optimal performance is achieved only when both RGB and Depth inputs are utilized, which yields a Collision Rate of 0.40 and a Success Rate of 0.23. This indicates that RGB data provides essential semantic context for the VLM and pose estimation, while Depth data ensures accurate trajectory forecasting and safety constraint enforcement.

(a) Instruction: *...Go through the corridor with the elevator, then walk to the two people by the green plants who appear to be reading and talking. Stop in front of them and wait for the next instruction...*

(b) Instruction: *...Go straight ahead for a short distance before the corridor, then turn right. Watch out for a person on the phone walking back and forth near a pillar. Go around him and move to the printer....*

(c) Instruction: *...Navigate across the open lounge towards the potted plant on the right. Watch out for the person walking past and safely avoid them...*

(d) Instruction: *...Move forward along the hallway to find the printer on your right, ensuring to safely bypass the person standing still in the middle of your path...*

(e) Instruction: *... Cross the corridor, minding any unseen people appearing at the corner...*

Fig. 7. Qualitative results of real-world deployment on the NXROBO Leo platform. (a) Semantic reasoning enables the agent to approach and wait for interacting pedestrians. (b) Geometric forecasting allows proactive detour planning around a pacing person. (c) The agent safely avoids a pedestrian crossing its path in an open lounge. (d) It correctly identifies and bypasses a stationary human blocking a hallway. (e) It successfully anticipates and yields to an unseen pedestrian appearing from a blind corner.

TABLE VIII
QUANTITATIVE RESULTS OF REAL-WORLD DEPLOYMENT. WE COMPARE THE SUCCESS RATE OF HCSG AGAINST BEVBERT ACROSS THE FIVE DISTINCT SCENARIOS DEPICTED IN FIG. 7, WITH 10 TRIALS PER SCENARIO.

Real-World Scenarios	Success Rate ↑	
	BEVBert [1]	HCSG (Ours)
Scenario (a): Stop in front of interacting pedestrians	3/10 (30%)	**9/10 (90%)**
Scenario (b): Avoid pacing pedestrian on phone	3/10 (30%)	**7/10 (70%)**
Scenario (c): Navigate to plant, avoid moving human	5/10 (50%)	**9/10 (90%)**
Scenario (d): Bypass standing human in hallway	6/10 (60%)	**9/10 (90%)**
Scenario (e): Traverse corridor, avoid corner pedestrian	2/10 (20%)	**7/10 (70%)**
Overall Average	19/50 (38%)	**41/50 (82%)**

D. Real-world Deployment

To validate the sim-to-real transferability of HCSG, we conducted physical experiments using the NXROBO Leo mobile manipulator equipped with an RGB-D camera in a controlled office environment. We evaluated the system through

both a systematic quantitative comparison and qualitative case demonstrations.

To systematically quantify the real-world capabilities, we established an evaluation benchmark comprising five distinct scenarios (depicted in Fig. 7) and conducted 10 physical trials per scenario. As summarized in Table VIII, HCSG provides preliminary real-world validation, achieving an overall success rate of 82%. In semantics-dependent tasks involving stationary humans (Scenarios a, d), BEVBert frequently failed by misclassifying interactive targets as mere obstacles or freezing, whereas our semantic stream ensured robust instruction grounding. Furthermore, in highly dynamic settings (Scenarios b, c, e), our geometric forecasting enabled proactive detour planning, yielding substantial improvements over the baseline's purely reactive approach and verifying the framework's effectiveness in real physical environments.

Beyond the quantitative metrics, Fig. 7 provides qualitative visualizations that highlight the mechanisms behind our agent's robustness. First, the framework demonstrates accurate semantic understanding when handling stationary individuals. For instance, in Fig. 7(a), the Semantic Reasoning stream successfully identifies two talking pedestrians as the interaction targets, navigating to a socially appropriate distance to wait. Similarly, in Fig. 7(d), the agent correctly parses a stationary human blocking a hallway as a passive obstacle to be bypassed rather than an interaction target.

Moreover, our agent exhibits superior geometric foresight when encountering dynamic obstacles. In Fig. 7(b), the Geometric Reasoning module captures the temporal motion of a pacing pedestrian near a pillar and proactively plans a detour. This proactive capability extends seamlessly to diverse spatial layouts: the robot intelligently avoids a pedestrian crossing an open lounge (Fig. 7(c)) and safely handles a dynamic encounter with an unseen pedestrian suddenly appearing from a blind corner (Fig. 7(e)) without freezing or causing collisions.

VI. Conclusion

In this work, we presented HCSG, a novel human-centric framework that endows VLN agents with explicit dual-stream human understanding. By combining geometric forecasting with semantic interpretation, our method enables the agent to reason about both future human motion and high-level human intent, moving beyond passive obstacle avoidance toward socially aware navigation. In addition, the proposed safety objective encourages the agent to maintain appropriate interaction distances around humans. Experimental results on the HA-VLNCE benchmark demonstrate that HCSG consistently improves both navigation success and collision reduction, highlighting the importance of explicit human-centric reasoning in dynamic indoor environments. Future work will focus on extending the current stop-and-wait design to a continuous streaming setting and on learning more adaptive social distance priors conditioned on richer human context. We believe these directions will further advance socially intelligent navigation in real-world human-robot environments.

References

[1] D. An, Y. Qi, Y. Li, Y. Huang, L. Wang, T. Tan, and J. Shao, "BEVBert: Multimodal map pre-training for language-guided navigation," *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023.

[2] 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.

[3] P. Anderson, Q. Wu, D. Teney, J. Bruce, M. Johnson, N. Sünderhauf, I. Reid, S. Gould, and A. 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*, 2018, pp. 3674–3683.

[4] A. Ku, P. Anderson, R. Patel, E. Ie, and J. Baldridge, "Room-Across-Room: Multilingual Vision-and-Language Navigation with Dense Spatiotemporal Grounding," in *Proc. Conf. Empirical Methods Nat. Lang. Process. (EMNLP)*, Nov. 2020, pp. 4392–4412.

[5] J. Krantz, E. Wijmans, A. Majumdar, D. Batra, and S. Lee, "Beyond the Nav-Graph: Vision-and-language navigation in continuous environments," in *European Conference on Computer Vision*. Springer, 2020, pp. 104–120.

[6] J. Chen, B. Lin, R. Xu, Z. Chai, X. Liang, and K.-Y. K. 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*, 2024.

[7] K. Su, X. Zhang, S. Zhang, J. Zhu, and B. Zhang, "To boost zero-shot generalization for embodied reasoning with vision-language pre-training," *IEEE Transactions on Image Processing*, vol. 33, pp. 5370–5381, 2024.

[8] G. Zhou, Y. Hong, and Q. Wu, "NavGPT: Explicit reasoning in vision-and-language navigation with large language models," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, no. 7, 2024, pp. 7641–7649.

[9] R. Quan, L. Zhu, Y. Wu, and Y. Yang, "Holistic lstm for pedestrian trajectory prediction," *IEEE transactions on image processing*, vol. 30, pp. 3229–3239, 2021.

[10] Y. Dong, F. Wu, Q. He, H. Li, M. Li, Z. Cheng, Y. Zhou, J. Sun, Q. Dai, Z.-Q. Cheng *et al.*, "HA-VLN: A benchmark for human-aware navigation in discrete-continuous environments with dynamic multi-human interactions, real-world validation, and an open leaderboard," *arXiv preprint arXiv:2503.14229*, 2025.

[11] A. Szot, A. Clegg, E. Undersander, E. Wijmans, Y. Zhao, J. Turner, N. Maestre, M. Mukadam, D. Chaplot, O. Maksymets, A. Gokaslan, V. Vondrus, S. Dharur, F. Meier, W. Galuba, A. Chang, Z. Kira, V. Koltun, J. Malik, M. Savva, and D. Batra, "Habitat 2.0: Training home assistants to rearrange their habitat," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

[12] M. Savva, A. Kadian, O. Maksymets, Y. Zhao, E. Wijmans, B. Jain, J. Straub, J. Liu, V. Koltun, J. Malik, D. Parikh, and D. Batra, "Habitat: A Platform for Embodied AI Research," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019.

[13] X. Puig, E. Undersander, A. Szot, M. D. Cote, R. Partsey, J. Yang, R. Desai, A. W. Clegg, M. Hlavac, T. Min, T. Gervet, V. Vondruš, V.-P. Berges, J. Turner, O. Maksymets, Z. Kira, M. Kalakrishnan, J. Malik, D. S. Chaplot, U. Jain, D. Batra, A. Rai, and R. Mottaghi, "Habitat 3.0: A Co-Habitat for Humans, Avatars and Robots," in *International Conference on Learning Representations*, 2024.

[14] S. Liu, A. Hasan, K. Hong, R. Wang, P. Chang, Z. Mizrachi, J. Lin, D. L. McPherson, W. A. Rogers, and K. Driggs-Campbell, "DRAGON: A dialogue-based robot for assistive navigation with visual language grounding," *IEEE Robotics and Automation Letters*, vol. 9, no. 4, pp. 3712–3719, 2024.

[15] J. Wang, E. B. Küçüktabak, R. S. Zarrin, and Z. Erickson, "CoRI: Communication of robot intent for physical human-robot interaction," in *9th Annual Conference on Robot Learning*, 2025.

[16] A. Payandeh, D. Song, M. Nazeri, J. Liang, P. Mukherjee, A. H. Raj, Y. Kong, D. Manocha, and X. Xiao, "Social-LLaVa: Enhancing robot navigation through human-language reasoning in social spaces," *arXiv preprint arXiv:2501.09024*, 2024.

[17] D. Song, J. Liang, A. Payandeh, A. H. Raj, X. Xiao, and D. Manocha, "VLM-Social-Nav: Socially aware robot navigation through scoring using vision-language models," *IEEE Robotics and Automation Letters*, vol. 10, no. 1, pp. 508–515, 2025.

[18] Y. Hong, Z. Wang, Q. Wu, and S. 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)*, 2022, pp. 15 418–15 428.

[19] D. An, Y. Qi, Y. Huang, Q. Wu, L. Wang, and T. Tan, "Neighbor-view enhanced model for vision and language navigation," in *Proceedings of the 29th ACM International Conference on Multimedia*, 2021, pp. 5101–5109.

[20] 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*, 2022, pp. 3617–3627.

[21] Z. He, L. Wang, S. Li, Q. Yan, C. Liu, and Q. Chen, "A multilevel attention network with sub-instructions for continuous vision-and-language navigation," *Applied Intelligence*, vol. 55, no. 7, 2025.

[22] 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*, 2021, pp. 1643–1653.

[23] X. Wang, Q. Huang, A. Celikyilmaz, J. Gao, D. Shen, Y.-F. Wang, W. Y. Wang, and L. Zhang, "Reinforced cross-modal matching and self-supervised imitation learning for vision-language navigation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 6629–6638.

[24] T. Li, W. Chen, H. Xu, X. Zheng, and H. Li, "P^{3}Nav: End-to-end perception, prediction and planning for vision-and-language navigation," *arXiv preprint arXiv:2603.17459*, 2026.

[25] J. Li, H. Tan, and M. Bansal, "EnvEdit: Environment editing for vision-and-language navigation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 15 407–15 417.

[26] Z. Wang, J. Li, Y. Hong, Y. Wang, Q. Wu, M. Bansal, S. Gould, H. Tan, and Y. Qiao, "Scaling data generation in vision-and-language navigation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 12 009–12 020.

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

[28] H. Tan, L. Yu, and M. Bansal, "Learning to Navigate Unseen Environments: Back Translation with Environmental Dropout," in *Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2019, pp. 2610–2621.

[29] H. Xu, T. Li, W. Chen, Y. Liu, X. Zuo, Y. Song, and H. Li, "Enhancing vision-language navigation with multimodal event knowledge from real-world indoor tour videos," 2026.

[30] P.-L. Guhur, M. Tapaswi, S. Chen, I. Laptev, and C. Schmid, "AirBERT: In-domain pretraining for vision-and-language navigation," in *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021, pp. 1614–1623.

[31] W. Hao, C. Li, X. Li, L. Carin, and J. Gao, "Towards learning a generic agent for vision-and-language navigation via pre-training," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 13 137–13 146.

[32] H. Huang, V. Jain, H. Mehta, A. Ku, G. Magalhaes, J. Baldridge, and E. Ie, "Transferable representation learning in vision-and-language navigation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 7404–7413.

[33] A. Majumdar, A. Shrivastava, S. Lee, P. Anderson, D. Parikh, and D. Batra, "Improving vision-and-language navigation with image-text pairs from the web," in *European Conference on Computer Vision*. Springer, 2020, pp. 259–274.

[34] Y. Qiao, Y. Qi, Y. Hong, Z. Yu, P. Wang, and Q. Wu, "HOP+: History-enhanced and order-aware pre-training for vision-and-language navigation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 7, pp. 8524–8537, 2023.

[35] G. Dai, S. Wang, H. Zhao, B. Zhu, Q. Sun, and X. Shu, "ThinkMatter: Panoramic-aware instructional semantics for monocular vision-and-language navigation," *IEEE Transactions on Image Processing*, 2026.

[36] W. Shi, C. Chen, K. Li, Y. Xiong, X. Cao, and Z. Zhou, "LangLoc: Language-driven localization via formatted spatial description generation," *IEEE Transactions on Image Processing*, 2025.

[37] S. Chen, P.-L. Guhur, M. Tapaswi, C. Schmid, and I. Laptev, "Think Global, Act Local: Dual-scale graph transformer for vision-and-language navigation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 16 537–16 547.

[38] 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*, 2023, pp. 15 625–15 636.

[39] R. Liu, W. Wang, and Y. Yang, "Volumetric environment representation for vision-language navigation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 16 317–16 328.

[40] K. Niu, L. Huang, Y. Long, Y. Huang, L. Wang, and Y. Zhang, "Comprehensive attribute prediction learning for person search by language," *IEEE Transactions on Image Processing*, vol. 33, pp. 1990–2003, 2024.

[41] K.-L. Wang, L.-W. Tsao, J.-C. Wu, H.-H. Shuai, and W.-H. Cheng, "TrajFine: Predicted trajectory refinement for pedestrian trajectory forecasting," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2024, pp. 4483–4492.

[42] G. Delmas, P. Weinzaepfel, T. Lucas, F. Moreno-Noguer, and G. Rogez, "PoseScript: 3d human poses from natural language," in *European Conference on Computer Vision*. Springer, 2022, pp. 346–362.

[43] L.-H. Chen, S. Lu, A. Zeng, H. Zhang, B. Wang, R. Zhang, and L. Zhang, "MotionLLM: Understanding human behaviors from human motions and videos," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 1–15, 2025.

[44] H. Li, M. Li, Z.-Q. Cheng, Y. Dong, Y. Zhou, J.-Y. He, Q. Dai, T. Mitamura, and A. G. Hauptmann, "Human-aware vision-and-language navigation: Bridging simulation to reality with dynamic human interactions," *Advances in Neural Information Processing Systems*, vol. 37, pp. 119 411–119 442, 2024.

[45] Z. Zhang, Z. Ding, and R. Tian, "Decouple ego-view motions for predicting pedestrian trajectory and intention," *IEEE Transactions on Image Processing*, vol. 33, pp. 4716–4727, 2024.

[46] A. H. Raj, Z. Hu, H. Karnan, R. Chandra, A. Payandeh, L. Mao, P. Stone, J. Biswas, and X. Xiao, "Rethinking Social Robot Navigation: Leveraging the best of two worlds," in *2024 IEEE International Conference on Robotics and Automation (ICRA)*, 2024, pp. 16 330–16 337.

[47] G. Pérez, N. Zapata-Cornejo, P. Bustos, and P. Núñez, "Social elastic band with prediction and anticipation: Enhancing real-time path trajectory optimization for socially aware robot navigation," *International Journal of Social Robotics*, vol. 17, no. 10, pp. 2041–2063, 2025.

[48] S. Samavi, J. R. Han, F. Shkurti, and A. P. Schoellig, "SICNav: Safe and interactive crowd navigation using model predictive control and bilevel optimization," *IEEE Transactions on Robotics*, vol. 41, p. 801–818, 2025.

[49] J. Li, J. He, W. Liu, T. Huang, S. Zhou, J. Ma, H. Wang, and H. Li, "SCSV: Spatial-temporal consistent dynamic 3d scene generation from sparse views," *IEEE Transactions on Image Processing*, 2026.

[50] J.-L. Bastarache, C. Nielsen, and S. L. Smith, "On legible and predictable robot navigation in multi-agent environments," in *2023 IEEE International Conference on Robotics and Automation (ICRA)*, 2023, pp. 5508–5514.

[51] B. Xue, M. Gao, C. Wang, Y. Cheng, and F. Zhou, "Crowd-aware socially compliant robot navigation via deep reinforcement learning," *International Journal of Social Robotics*, vol. 16, no. 1, pp. 197–209, 2024.

[52] Z. Sun, X. Diao, Y. Wang, B.-K. Zhu, and J. Wang, "Socially aware robot crowd navigation via online uncertainty-driven risk adaptation," *arXiv preprint arXiv:2506.14305*, 2025.

[53] Z. Gong, T. Hu, R. Qiu, and J. Liang, "From cognition to precognition: A future-aware framework for social navigation," in *2025 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2025, pp. 9122–9129.

[54] R. Hartley and A. Zisserman, *Multiple view geometry in computer vision*. Cambridge university press, 2003.

[55] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, "Learning transferable visual models from natural language supervision," in *International conference on machine learning*. PmLR, 2021, pp. 8748–8763.

[56] S. Raychaudhuri, S. Wani, S. Patel, U. Jain, and A. Chang, "Language-aligned waypoint (law) supervision for vision-and-language navigation in continuous environments," in *Proceedings of the 2021 conference on empirical methods in natural language processing*, 2021, pp. 4018–4028.

[57] D. Maji, S. Nagori, M. Mathew, and D. Poddar, "YOLO-Pose: Enhancing yolo for multi person pose estimation using object keypoint similarity loss," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 2637–2646.

[58] Q. Team, "Qwen3 technical report," 2025. [Online]. Available: https://arxiv.org/abs/2505.09388