FreeAskWorld: An Interactive and Closed-Loop Simulator for Human-Centric Embodied AI

Yuhang Peng, Yizhou Pan, Xinning He, Jihaoyu Yang, Xinyu Yin, Han Wang, Xiaoji Zheng, Chao Gao, Jiangtao Gong

Institute for AI Industry Research, Tsinghua University
gongjiangtao@air.tsinghua.edu.cn

Abstract

As embodied intelligence emerges as a core frontier in artificial intelligence research, simulation platforms must evolve beyond low-level physical interactions to capture complex, human-centered social behaviors. We introduce FreeAskWorld, an interactive simulation framework that integrates large language models (LLMs) for high-level behavior planning and semantically grounded interaction, informed by theories of intention and social cognition. Our framework supports scalable, realistic human-agent simulations and includes a modular data generation pipeline tailored for diverse embodied tasks. To validate the framework, we extend the classic Vision-and-Language Navigation (VLN) task into a interaction enriched Direction Inquiry setting, wherein agents can actively seek and interpret navigational guidance. We present and publicly release FreeAskWorld, a large-scale benchmark dataset comprising reconstructed environments, six diverse task types, 16 core object categories, 63,429 annotated sample frames, and more than 17 hours of interaction data to support training and evaluation of embodied AI systems. We benchmark VLN models and human participants under both open-loop and closed-loop settings. Experimental results demonstrate that models fine-tuned on FreeAskWorld outperform their original counterparts, achieving enhanced semantic understanding and interaction competency. These findings underscore the efficacy of socially grounded simulation frameworks in advancing embodied AI systems toward sophisticated high-level planning and more naturalistic human-robot interaction.

Code — https://github.com/AIR-DISCOVER/FreeAskWorld

Datasets — https://huggingface.co/datasets/Astronaut-PENG/FreeAskWorld

Introduction

Understanding and following human-generated navigation instructions is a critical capability for embodied AI agents operating in real-world environments. Vision-and-Language Navigation (VLN) tasks, which require agents to interpret natural language directives and traverse complex visual scenes, have emerged as a central research area integrating computer vision, natural language processing, and robotics. Benchmarks such as Room-to-Room (R2R) and its variants have significantly advanced progress in this domain. However, existing VLN systems are still constrained by three core limitations.

First, most methods rely on static, one-shot instructions provided at the beginning of a navigation episode, limiting the agent's ability to handle dynamic goals or engage in multi-turn interactions. Second, current VLN frameworks often decouple high-level planning from social intention modeling, resulting in agents that cannot interpret socially salient cues or perform context-aware, human-like behaviors. Third, despite improvements in realism, simulators supporting VLN research—such as Grutopia—frequently lack complex, interactive, and dynamic elements such as moving pedestrians, social interactions, and real-world environmental variation, making them insufficient for modeling socially grounded human-agent communication.

Simultaneously, the development of generative AI and large language models has opened new possibilities for modeling high-level behaviors in simulation environments. Recent works have demonstrated that LLM-driven agents can generate diverse goals, simulate social interactions, and perform role-based tasks in virtual societies. Nevertheless, these approaches often remain disconnected from embodied navigation research, and lack mechanisms for real-time closed-loop interaction grounded in semantic and spatial understanding.

To bridge this gap, we propose FreeAskWorld—an interactive and closed-loop simulation framework designed for human-centric embodied AI. Our system leverages large language models for high-level intention modeling, semantic instruction generation, and naturalistic human behavior simulation. Grounded in theories of social behavior and intention, FreeAskWorld enables dynamic and contextually rich interactions between AI agents and human-like avatars within photorealistic 3D environments.

We further extend the classical VLN paradigm by introducing the Direction Inquiry Task, a novel benchmark that allows agents to proactively seek help and adjust their navigation based on new information, thereby evaluating higher-order capabilities such as self-assessment, social interaction, and real-time adaptation. To support this task, we construct and release the FreeAskWorld Dataset, a large-scale synthetic dataset featuring diverse human avatars, reconstructed

urban environments, dynamic vehicles, and six categories of synthetic data (e.g., dialog histories, panoramic RGB, occupancy maps). The dataset includes over six hours of interactive simulation data and supports both open-loop and closed-loop evaluation.

Our work aims to fill this gap by introducing:

- An interactive, LLM-driven simulation framework featuring dynamic and realistic human agents with high-level intention modeling, semantic interaction control, and a synthetic data generation pipeline.
- A novel benchmark task for direction inquiries that enables controllable and extensible devaluation of human-centric social navigation and interaction.
- We evaluated several representative VLN models, along with a human baseline, to validate the rationality of the proposed task and simulator, as well as the effectiveness of the generated data.

Related Work

Simulating Human Behavior in Social Navigation

Core behavioral challenges in social navigation span from individual-level proxemics to macro-level social signaling in multi-agent contexts (Rios-Martinez, Spalanzani, and Laugier 2015; Mavrogiannis et al. 2023). Traditional simulation platforms have primarily relied on low-level physical behavior models to simulate crowd dynamics and agent-level reactions to robots (Grzeskowiak et al. 2021; Tsoi et al. 2022; Vuong et al. 2024; Pérez-Higueras et al. 2023). Although effective in modeling local interactions and physical constraints, these approaches often miss the dynamic and contextual nuances of human behavior at the societal-level. To address this, recent work has turned to generative AI to simulate high-level, context-aware behaviors(Wang et al. 2024b; Xi et al. 2025; Piao et al. 2025). For instance, MARPLE introduces a hierarchical structure that decomposes behavior into missions, subgoals, and atomic actions (Jin et al. 2024). Recent simulators, such as Virtual Community(Zhou et al. 2025), MetaUrban(Wu et al. 2024), and Grutopia(Wang et al. 2024a), push this further to urban-scale world by modeling open-ended social interactions among autonomous generative agents.

Building on these advances, we propose a simulation framework that integrates LLMs for high-level planning and semantically grounded interaction, informed by theories of intention and social behavior. Combined with Unity's animation engine and asset library, our approach enables realistic, scalable, and interactive human-centric simulations.

Human-Centric Language Use in Navigation

Wayfinding through verbal directions is a complex and dynamic process, shaped by the direction giver, the environment, the task at hand, and the recipient(Hund, Haney, and Seanor 2008). Two common strategies in verbal instructions are the route perspective, using egocentric cues like landmarks and left or right turns; and the survey perspective, based on map-like descriptions involving distances, street names, and cardinal directions. The navigation style from the perspective of the direction giver is mainly influenced by socio-spatial background and gender (Galea and Kimura 1993; Lawton 1996, 2001; Kato and Takeuchi 2003). For example, irregular European layouts tend to favor landmark-based navigation, while grid-based American cities rely more on street names(Hund, Schmettow, and Noordzij 2012). Women usually use more spatial references, provide longer instructions, and incorporate hedging expressions(Sing and Kalingga 2011).

Vision Language Navigation

Vision-and-Language Navigation (VLN) is a foundational problem in embodied AI, where an agent interprets natural language instructions to navigate visually rich 3D environments. This interdisciplinary task integrates computer vision, natural language understanding, and robotics, driving progress toward agents capable of understanding and following human-like directions in realistic settings.

The classic VLN task was introduced by the Room-to-Room benchmark (Anderson et al. 2018), which provides natural language navigation instructions grounded in real-world indoor panoramic scenes. However, R2R relies on discrete navigation graphs, limiting realism and fine-grained spatial interaction. REVERIE (Qi et al. 2020) extends R2R by introducing object grounding, but still lacks continuous motion and physical embodiment. R2R-CE (Krantz et al. 2020) addresses this by enabling continuous action spaces with real-time physics, introducing challenges like low-level control and precise localization. Talk2Nav (Vasudevan, Dai, and Van Gool 2021) further advances VLN with long-range instruction following using dual attention mechanisms and spatial memory modules.

Despite progress, VLN systems face three key limitations: (1) reliance on static one-shot instructions, reducing adaptability to dynamic environments and multi-turn interactions; (2) lack of integration between high-level planning and social intention modeling, limiting agents' ability to interpret social signals and perform context-aware behaviors; and (3) simulators that do not capture the complexity of real-world environments with dynamic human agents, moving vehicles, and socially grounded interactions (Wang et al. 2024a).

To address these gaps, we propose a socially aware VLN framework that integrates realistic simulation environments with dynamic, human-like instruction generation. Using large language models and behavior priors, our system enables real-time socially grounded interactions, unifying language understanding, motor control, and social reasoning in a continuous embodied setting.

FreeAskWorld: Simulator Design

Motivation and Overview

To facilitate the simulation of high-level human-computer interaction tasks, it is essential to achieve high-fidelity modeling of human behavior as well as complex social systems. Existing simulation environments primarily focus on physical-level or task-level interactions, often lacking realistic representations of social behavior, environmental dynamics, and long-term temporal evolution. This limitation

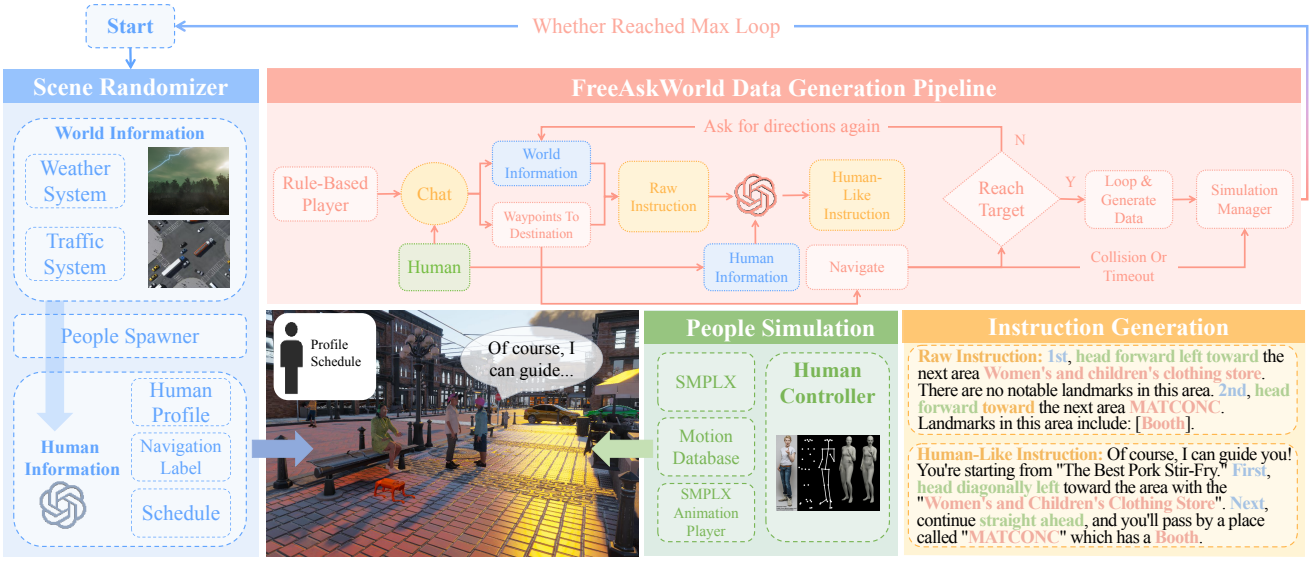

Figure 1: An overview of the FreeAskWorld framework and its data generation pipeline is presented. The system incorporates scene randomization techniques to enhance environmental diversity. The data generation module produces the FreeAskWorld dataset through this pipeline. People simulation module for modeling realistic human behaviors within virtual scenes, as well as an instruction generation module for producing navigation directives used in the Direction Inquiry Task within the simulator.

hinders the development and evaluation of advanced interaction strategies.

To address this gap, we propose FreeAskWorld, a high-fidelity simulation platform designed to model human-like societies with structured social behavior, open interaction interfaces, and integrated multisystem dynamics. FreeAskWorld incorporates real-world social structures and behavioral norms to enable multilevel behavior modeling and decision-making processes. In addition, the platform integrates several subsystems, such as transportation, weather, and daily activity cycles, to introduce environmental variability and unpredictability, thus enhancing the realism of the simulated world.

This design lays the foundation for evaluating a wide range of high-level interaction strategies in environments that closely approximate real-world social contexts. In the current stage, we focus on direction inquiries as a representative interaction task. However, FreeAskWorld is designed to support, in future developments, more complex interactions such as natural language negotiation, task coordination, social navigation, and long-term trust building.

Furthermore, FreeAskWorld supports large-scale multi-agent deployment and behavior tracking. Through a modular interface architecture, it enables seamless integration of language models, multimodal perception modules, and other intelligent components, offering a flexible and extensible platform for studying sustained interaction and autonomous learning in human-like societies.

People Simulation

This section introduces the People Simulation module, which encompasses avatar modeling, profile and schedule generation, navigation style synthesis, animation control, and appearance variation. An overview of the module is illustrated in Figure 3.

Avatar Models The visual and kinematic realism of human agents significantly impacts the credibility of Sim2Real navigation systems. Many navigation simulators still rely on static 3D character models with minimal animation diversity(Li et al. 2021; Vuong et al. 2024). HA-VLN takes a step forward by combining SMPL-based body models with AI-based action planning, enhancing the naturalness of character motion (Dong et al. 2025).

Profile And Schedule Generation We employ a two-stage generation framework to create diverse and realistic human agents. First, character profiles are synthesized incorporating demographic and contextual attributes, such as age, culture, and occupation. In the second stage, given the character profile and a static scene layout, we generate a corresponding daily schedule, which includes multiple activities with temporal segmentation and location assignments sampled from available destinations within the environment.

Navigation Style Generation We incorporate regional familiarity and personality into the role profile and classify navigation style by four key features: landmark use, direction type, distance description, and utterance length. The findings of the literature are encoded as a knowledge base, enabling the LLM to generate contextually grounded navigation labels and instructions based on the role profile, task, and geological situation.

Animation Database To support consistent embodiment and context-sensitive behaviors, we adopt MotionX(Lin et al. 2023) as the SMPL-X animation library and structure

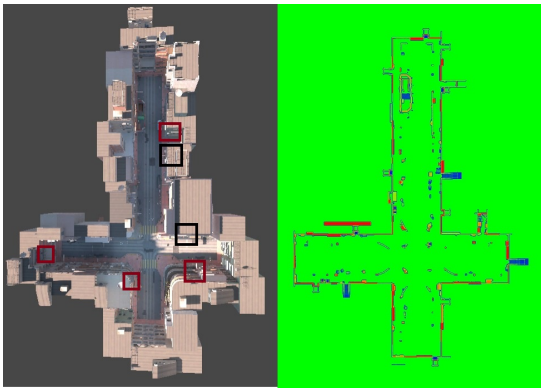

Figure 2: Comparison Between the Original Mesh Model and the Generated Occupancy Heatmap. The black and red bounding signs represent the same store A and B in different positions, a layout designed to assess the navigation capabilities of humans or robots in complex environments.

its motions into high-level categories and subcategories, allowing the selection of semantically relevant motions for each activity. We also developed a custom SMPL-X animation driver plugin that allows seamless access to the full range of animations within the library. To enhance animation fidelity and character realism, we integrated Blending functionality into the motion controller, enabling smooth transitions between actions.

Apperance Variation *Method 1:* We propose a new framework that uses Multimodal Large Language Models (MLLMs) to create diverse virtual human appearances controlled by natural language. By combining UV mapping with semantic profiles (such as gender, occupation, and ethnicity), we build a comprehensive pipeline for texture generation guided by language, enabling scalable and consistent material creation across a variety of scenarios. At the mesh level, we introduce variation by randomly modifying the shape parameters of the SMPL-X model, producing diverse body types characterized by differences in height, weight, and proportions. This approach, coupled with semantic control over beta shapes properties, enhances both the realism and expressiveness of human simulations. *Method 2:* We leverage the Synbody (Yang et al. 2023) dataset to generate SMPL-X models, which incorporate clothing, footwear, hair, and other detailed features, offering a more immersive and realistic virtual experience. This approach enhances the fidelity of modeling human appearance.

Other System Functions

Occupancy Map Generation We generate a 2D occupancy heat map based on voxels to support navigation, mapping, and interaction. The environment is divided into 3D voxels within a set region. For each voxel, occupancy is estimated probabilistically through random sampling and collision checks, producing a soft occupancy value.

To improve reliability, we repeat the sampling multiple times and average the results. The occupancy map is then refined using filtering, morphological operations, and noise removal techniques. Finally, the 3D map is projected onto a 2D plane to create top-down occupancy maps, which can be exported as heatmaps or binary grids.

An example is shown in Fig. 2.

Weather System We use a dynamic sky and weather system to simulate day-night cycles and various weather conditions like rain and fog. This improves simulation realism and data diversity, helping models trained on this data generalize better under different lighting and visibility.

Traffic Simulation We integrate a traffic simulation system that models vehicle movement and traffic rules using route graphs. This adds realistic vehicle behaviors and dynamic complexity to the environment, supporting more challenging perception and planning tasks.

Robot Simulation Our system uses the A* algorithm for global path planning and the Social Force Model (SFM) for local obstacle avoidance, enabling socially-aware navigation in dynamic environments. The application of SFM during data collection significantly improves navigation success rates by effectively avoiding pedestrians and vehicles, though it results in longer trajectories due to more cautious path planning. Furthermore, the simulator supports realistic robot dynamics through Unity's articulation components and allows seamless integration of physical robots via the URDF-Importer.

Synchronous Closed-loop Framework Considering that non-headless simulators typically cannot be run in server environments, we have designed and implemented a WebSocket-based synchronous closed-loop simulation architecture. This architecture supports closed-loop simulation through network connections between the server-side model and the simulator via NAT traversal techniques, as well as data exchange via port communication on the same device. The system provides diverse message interfaces to accommodate the transmission of various types of sensor data and control commands.

Direction Inquiry Task

We extend the traditional Vision-and-Language Navigation (VLN) task by introducing an inquiry phase, allowing the agent to actively seek external information. This modification enables the evaluation of the model's high-level capabilities, such as self-assessment, information-seeking behavior, and planning based on acquired knowledge. Additionally, our simulation environment includes a large number of dynamic humans and vehicles, facilitating the assessment of the model's low-level motion planning and control capabilities under realistic and dynamic conditions.

Evaluation Metrics

To comprehensively evaluate the agent's performance in the Direction Inquiry Task, we employ the following standard metrics commonly used in navigation and vision-language tasks:

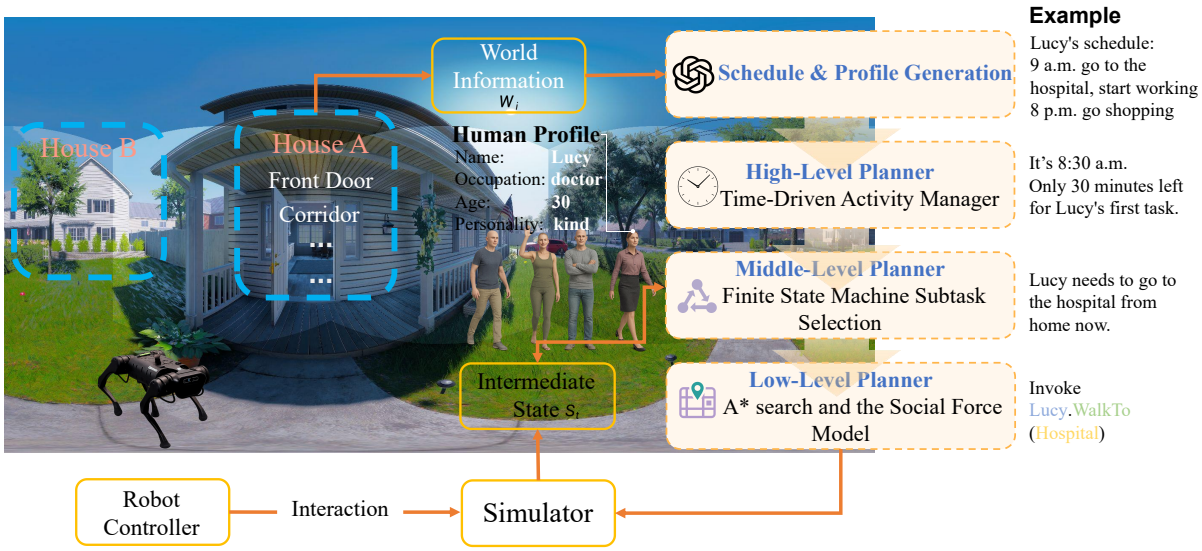

Figure 3: The People Simulation works as follows: a large language model creates diverse character profiles and schedules, which the high-level planner uses to select activities based on time. These activities are passed to the middle-level planner, which breaks them into subtasks and manages them with a finite state machine. Low-level planners handle navigation and basic behaviors, using A* for global paths and the Social Force Model to avoid obstacles locally.

- **Trajectory Length (TL)**: The average total distance traveled by the agent, computed as $\text{TL} = \frac{1}{N}\sum_{i=1}^{N}\sum_{t=1}^{T_i-1} d(p_i^{t+1}, p_i^t)$, where T_i is the number of time steps in episode i.
- **Success weighted by Path Length (SPL)**: Measures both success and efficiency (Anderson et al. 2018), defined as $\text{SPL} = \frac{1}{N}\sum_{i=1}^{N} S_i \frac{l_i^*}{\max(l_i, l_i^*)}$, where $S_i \in \{0, 1\}$ indicates success in episode i, l_i^* is the shortest path length to the goal, and l_i is the actual path length taken by the agent.
- **Success Rate (SR)**: The proportion of successful episodes, $\text{SR} = \frac{1}{N}\sum_{i=1}^{N} S_i$, where success is defined as reaching within a threshold distance δ of the goal.
- **Navigation Error (NE)**: The average final distance between the agent and the goal, $\text{NE} = \frac{1}{N}\sum_{i=1}^{N} d(p_i^{T_i}, g_i)$.
- **Oracle Navigation Error (ONE)**: The average minimum distance to the goal along the trajectory, $\text{ONE} = \frac{1}{N}\sum_{i=1}^{N} \min_{t=1,\ldots,T_i} d(p_i^t, g_i)$.
- **Oracle Success Rate (OSR)**: The fraction of episodes in which the agent is within δ of the goal at any point during its trajectory, $\text{OSR} = \frac{1}{N}\sum_{i=1}^{N} \mathbf{1}(\min_{t=1,\ldots,T_i} d(p_i^t, g_i) \le \delta)$, where δ is the radius of success (typically 1–3 m depending on the environment).
- **Number of Direction Inquiries (NDI)**: The average number of inquiry actions issued by the agent per episode, $\text{NDI} = \frac{1}{N}\sum_{i=1}^{N} n_i^{\text{inquiry}}$.

Dataset	Scene	Type	Inst. (words)	Traj. (m)
Talk2Nav	Outdoor	Discrete	∼70	∼60
R2R	Indoor	Discrete	∼29	∼20
REVERIE	Indoor	Discrete	∼18	∼12
ScaleVLN	Indoor	Discrete	∼29	∼10
NavRAG	Indoor	Discrete	∼25	∼8
GSA-R2R	Indoor	Discrete	∼35	∼12
HA-R2R	Indoor	Discrete	∼29	∼10
NaVid	Indoor	Continuous	∼20	∼25
DynamicVLN	Outdoor	Continuous	∼28	∼400
VLN-Video	Outdoor	Continuous	∼90	∼200
FreeAskWorld	In/Outdoor	Continuous	∼148	∼56

Table 1: Comparison of Vision-and-Language Navigation Datasets

FreeAskWorld Dataset

Dataset Overview

We compare our dataset with other existing VLN datasets, including Talk2Nav (Vasudevan, Dai, and Van Gool 2021), R2R (Anderson et al. 2018), REVERIE (Qi et al. 2020), ScaleVLN (Wang et al. 2023), NavRAG (Wang et al. 2025), GSA-R2R (Li et al. 2024a), HA-R2R (Li et al. 2024a), NaVid (Zhang et al. 2024), DynamicVLN (Sun, Qiu, and Aoki 2025), and VLN-Video (Li et al. 2024b). Table 1 summarizes their characteristics.

Dataset Validation

To validate the data in our dataset, we conduct a human baseline evaluation in Figure. 2. In this process, Experimenters

verify the alignment between the LLM-generated instructions and the intended destination. The generated text should exhibit qualities of being coherent, human-like, and easy to understand. This approach ensures that the instructions are both reasonable and effective in guiding the navigation process.

Data Generation Pipeline

The data generation pipeline begins with the initialization of the simulation environment, including randomization of environmental conditions such as weather and time of day to promote diversity and robustness. Once initialized, the data collection agent actively searches for nearby human agents and initiates an interaction to request navigational assistance. The responses are generated by a large language model (LLM) to simulate realistic, human-like instructions.

Following the dialogue, the agent navigates to the specified destination using a socially compliant navigation strategy that accounts for both static and dynamic obstacles. If the agent fails to reach the destination within a predefined time threshold, it will initiate another round of inquiry with nearby human agents to update its goal. Upon successful arrival at the target location, the episode is marked as successful, and all relevant data—including dialogue transcripts, panoramic and perspective RGB images, and associated synthetic annotations—are recorded for training and evaluation purposes

Sensor Configuration

We configured six cameras, each with a 90 ° field of view (FOV), positioned at the same spatial location but oriented in different directions to collectively capture panoramic images at a frequency of 1 Hz. During the data acquisition process, we simultaneously recorded the position and orientation of the camera rig in the world coordinate system. This information enables the transformation and alignment between the synthesized image data and the global scene context.

Trajectory Generation

In the simulated environment, we randomly assign start and target positions, followed by an inquiry-based behavior to obtain navigational guidance. The agent then navigates to the goal using a rule-based path planning algorithm, which accounts for static and dynamic obstacle avoidance in a socially compliant manner. The resulting trajectories serve as expert demonstrations for imitation learning.

Data Composition

We used Unity Perception(Borkman et al. 2021) to construct a rich and diverse synthetic dataset that encompasses multiple types of annotations and data modalities. The dataset is designed to support a wide range of vision and navigation tasks and includes both dense per-frame annotations and global scene-level metadata. Specifically, the dataset comprises:

- **Visual annotations**: 2D/3D bounding boxes, instance and semantic segmentation.

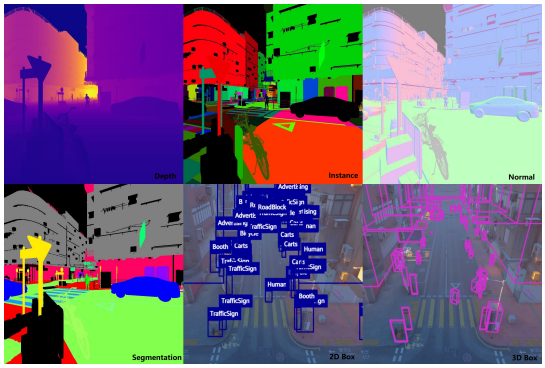

Figure 4: Six Main Types of Synthetic Data

- **Geometric annotations**: depth maps and surface normal maps for scene geometry.
- **Visual observations**: panoramic RGB images and six 90° perspective views.
- **Interaction data**: natural language instructions, dialog histories, and agent trajectories.
- **Spatial representations**: 2D occupancy heatmaps for mapping and localization.
- **Environment metadata**: map boundaries, semantic regions, and other contextual information.

As shown in Figure 4, the six main types of synthetic data are summarized.

Additionally, the dataset includes annotations for 16 key object categories commonly encountered in human-centered environments, such as vehicles, pedestrians, and street furniture.

Scene Reconstruction

Based on the 2D occupancy heatmaps that encode the layout of static elements, along with the 3D bounding boxes that capture the positions of dynamic entities, the simulation environment can be accurately reconstructed. By further integrating the provided world information, it becomes possible to generate a comprehensive digital twin of the scene. This reconstructed environment enables open-loop evaluations similar to those in the nuScenes dataset (Caesar et al. 2020), and is particularly well suited for unstructured settings as in FreeAD (Peng et al. 2025), and supports a broad spectrum of downstream tasks including navigation planning, behavior prediction, and human-computer interaction studies.

Experiments

We conduct model training and fine-tuning on the FreeAskWorld dataset. For evaluation, open-loop experiments are performed in the OpenAskWorld open-loop test split, while closed-loop experiments are conducted in both the closed-loop test split and within the simulator environment.These experiments serve to validate the effectiveness of our dataset and simulator.

Baselines

We conduct both open-loop and closed-loop evaluations on the FreeAskWorld dataset to comprehensively assess model performance.

We compare several baselines as follows:

- Human: Used as an upper bound reference for navigation performance.
- ETPNav (An et al. 2024): A hierarchical VLN-CE framework that performs online topological mapping, cross-modal planning with transformers, and low-level control using a rotate-then-forward schema augmented by obstacle-avoidance heuristics.
- BEVBert (An et al. 2022): A map-based multimodal pretraining model that leverages hybrid topo-metric representations to improve spatial reasoning and language-guided navigation robustness.

We evaluate ETPNav and BEVBert using both their original pretrained models and fine-tuned versions on the FreeAskWorld dataset, referred to as ETPNav-FT and BEVBert-FT.

Experimental Setup

For open-loop evaluation, models are tested on the FreeAskWorld open-loop test set. In closed-loop evaluation, each episode starts with scene conditions, agent pose, and the first navigation instruction, which does not count toward the Number of Direction Inquiries (NDI), starting at NDI = 0. Multiple trials per episode ensure robustness, with averaged results reported. The human baseline involves four participants: two follow the initial instruction only, while the other two may ask follow-up questions (each counted as one NDI). Participants control the agent with keyboard inputs. For model-based methods (ETPNav, ETPNav-FT, BEVBert, BEVBert-FT), we use a synchronous closed-loop framework with a 1 Hz update rate. Models run on an RTX 3080, and the simulator on an RTX 3060. Episodes terminate after pedestrian/vehicle collisions or after 100 steps.

Results

The open-loop results show that the fine-tuned models, `ETPNav-FT` and `BEVBert-FT`, achieve a $\sim$50% reduction in L2 error compared to their base versions, with `BEVBert-FT` delivering the best overall performance.

In closed-loop experiments, the human baseline demonstrates that agents capable of querying for additional navigation instructions significantly improve pathfinding accuracy (from 40.2% to 82.6%). This improvement is attributed to scene complexity, such as identical stores appearing in different directions within 50 meters (shown in Fig. 4), where humans may occasionally experience shallow memory of directions and become disoriented. Fine-tuned models, `ETPNav-FT` and `BEVBert-FT`, show substantial improvements over their base counterparts in both Navigation Error and Oracle Navigation Error. The increase in Trajectory Length suggests enhanced familiarity with the scene and broader exploration. Although `ETPNav-FT` achieves partial destination success, overcoming the zero Oracle Success Rate observed in baseline models, its overall Success Rate remains zero. A similar trend is observed in the Social Mobile Manipulation task of InfiniteWorld (Ren et al. 2024), where tasks involving social interactions substantially reduce robot performance. This challenge arises from weak dynamic social navigation, pedestrian/vehicle collisions, and limitations in long-range planning, abstract reasoning, memory retention, and higher-level decision-making—areas that warrant further investigation. Additionally, `BEVBert` consistently outperforms `ETPNav` across both evaluation metrics, reinforcing its state-of-the-art performance in Vision-and-Language Navigation (VLN) and highlighting commonalities with our task.

Method	TL	SR↑	SPL↑	NE↓	OSR↑	ONE↓	NDI
Human(no ask)	47.5	40.2	38.2	18.3	41.3	11.3	0.0
Human(ask)	59.9	**82.6**	**71.2**	**3.49**	**82.6**	**1.63**	0.78
ETPNav	31.2	0.0	0.0	32.9	0.0	28.7	0.0
BEVBert	14.6	0.0	0.0	31.0	0.0	29.0	0.0
ETPNav-FT	33.6	0.0	0.0	31.6	**1.1**	**27.1**	0.0
BEVBert-FT	18.7	0.0	0.0	**30.0**	0.0	28.5	0.0

Table 2: Closed-Loop Navigation Performance of Various Methods in the FreeAskWorld Simulator

Overall, these evaluations not only validate the effectiveness of the proposed dataset but also shed light on the strengths and limitations of current models in socially situated navigation as well as in the Direction Inquiry Task or Interaction-oriented tasks.

Conclusion

We introduce the Direction Inquiry Task to extend traditional VLN settings, emphasizing self-assessment, social interaction, and real-time adaptation. To support this, we release the FreeAskWorld Dataset, featuring diverse avatars, dynamic scenes, and rich multimodal annotations. Experiments show that VLN models fine-tuned on our dataset improve in both open-loop and closed-loop settings. However, comparisons with human performance reveal that current models still struggle with high-level reasoning and socially grounded navigation.

Importantly, our work underscores that **interaction itself serves as an additional information modality**. Intentional and structured interaction is **not only a social signal but also a crucial pathway for understanding and interpreting the physical world**, enabling agents to acquire information that static perception alone cannot provide. This highlights the broader value of socially grounded simulation in bridging the gap between embodied AI and real-world human interaction.

Future directions include tackling complex tasks like negotiation and coordination, integrating multimodal memory and perception for adaptive behaviors, and leveraging generative models for higher visual fidelity. We also plan to develop an end-to-end software solution on Steam for easier access and expand the benchmark suite to include more comprehensive metrics to evaluate embodied AI interactions.

Acknowledgments

This work was supported by Beijing Natural Science Foundation L233033 and China Natural Science Foundation Youth Fund 62202267.

References

An, D.; Qi, Y.; Li, Y.; Huang, Y.; Wang, L.; Tan, T.; and Shao, J. 2022. Bevbert: Multimodal map pre-training for language-guided navigation. *arXiv preprint arXiv:2212.04385*.

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

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

Borkman, S.; Crespi, A.; Dhakad, S.; Ganguly, S.; Hogins, J.; Jhang, Y.-C.; Kamalzadeh, M.; Li, B.; Leal, S.; Parisi, P.; et al. 2021. Unity perception: generate synthetic data for computer vision. *arXiv preprint arXiv:2107.04259*.

Caesar, H.; Bankiti, V.; Lang, A. H.; Vora, S.; Liong, V. E.; Xu, Q.; Krishnan, A.; Pan, Y.; Baldan, G.; and Beijbom, O. 2020. nuscenes: A multimodal dataset for autonomous driving. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 11621–11631.

Dong, Y.; Wu, F.; He, Q.; Li, H.; Li, M.; Cheng, Z.; Zhou, Y.; Sun, J.; Dai, Q.; Cheng, Z.-Q.; et al. 2025. 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*.

Galea, L. A.; and Kimura, D. 1993. Sex differences in route-learning. *Personality and individual differences*, 14(1): 53–65.

Grzeskowiak, F.; Gonon, D.; Dugas, D.; Paez-Granados, D.; Chung, J. J.; Nieto, J.; Siegwart, R.; Billard, A.; Babel, M.; and Pettré, J. 2021. Crowd against the machine: A simulation-based benchmark tool to evaluate and compare robot capabilities to navigate a human crowd. In *2021 IEEE International Conference on Robotics and Automation (ICRA)*, 3879–3885. IEEE.

Hund, A. M.; Haney, K. H.; and Seanor, B. D. 2008. The role of recipient perspective in giving and following wayfinding directions. *Applied Cognitive Psychology: The Official Journal of the Society for Applied Research in Memory and Cognition*, 22(7): 896–916.

Hund, A. M.; Schmettow, M.; and Noordzij, M. L. 2012. The impact of culture and recipient perspective on direction giving in the service of wayfinding. *Journal of Environmental Psychology*, 32(4): 327–336.

Jin, E.; Huang, Z.; Fränken, J.-P.; Liu, W.; Cha, H.; Brockbank, E.; Wu, S.; Zhang, R.; Wu, J.; and Gerstenberg, T. 2024. MARPLE: A benchmark for long-horizon inference. *Advances in Neural Information Processing Systems*, 37: 108824–108850.

Kato, Y.; and Takeuchi, Y. 2003. Individual differences in wayfinding strategies. *Journal of environmental psychology*, 23(2): 171–188.

Krantz, J.; Wijmans, E.; Majumdar, A.; Batra, D.; and Lee, S. 2020. Beyond the nav-graph: Vision-and-language navigation in continuous environments. In *European Conference on Computer Vision*, 104–120. Springer.

Lawton, C. A. 1996. Strategies for indoor wayfinding: The role of orientation. *Journal of environmental psychology*, 16(2): 137–145.

Lawton, C. A. 2001. Gender and regional differences in spatial referents used in direction giving. *Sex Roles*, 44(5): 321–337.

Li, C.; Xia, F.; Martín-Martín, R.; Lingelbach, M.; Srivastava, S.; Shen, B.; Vainio, K.; Gokmen, C.; Dharan, G.; Jain, T.; et al. 2021. igibson 2.0: Object-centric simulation for robot learning of everyday household tasks. *arXiv preprint arXiv:2108.03272*.

Li, H.; Li, M.; Cheng, Z.-Q.; Dong, Y.; Zhou, Y.; He, J.-Y.; Dai, Q.; Mitamura, T.; and Hauptmann, A. G. 2024a. Human-aware vision-and-language navigation: Bridging simulation to reality with dynamic human interactions. *Advances in Neural Information Processing Systems*, 37: 119411–119442.

Li, J.; Padmakumar, A.; Sukhatme, G.; and Bansal, M. 2024b. Vln-video: Utilizing driving videos for outdoor vision-and-language navigation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 18517–18526.

Lin, J.; Zeng, A.; Lu, S.; Cai, Y.; Zhang, R.; Wang, H.; and Zhang, L. 2023. Motion-X: A Large-scale 3D Expressive Whole-body Human Motion Dataset. *Advances in Neural Information Processing Systems*.

Mavrogiannis, C.; Baldini, F.; Wang, A.; Zhao, D.; Trautman, P.; Steinfeld, A.; and Oh, J. 2023. Core Challenges of Social Robot Navigation: A Survey. *J. Hum.-Robot Interact.*, 12(3).

Peng, Y.; Wang, S.; Yang, J.; Li, S.; Wang, H.; and Gong, J. 2025. Bench2FreeAD: A Benchmark for Vision-based End-to-end Navigation in Unstructured Robotic Environments. *arXiv preprint arXiv:2503.12180*.

Pérez-Higueras, N.; Otero, R.; Caballero, F.; and Merino, L. 2023. Hunavsim: A ros 2 human navigation simulator for benchmarking human-aware robot navigation. *IEEE robotics and automation letters*, 8(11): 7130–7137.

Piao, J.; Yan, Y.; Zhang, J.; Li, N.; Yan, J.; Lan, X.; Lu, Z.; Zheng, Z.; Wang, J. Y.; Zhou, D.; et al. 2025. Agentsociety: Large-scale simulation of llm-driven generative agents advances understanding of human behaviors and society. *arXiv preprint arXiv:2502.08691*.

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

Ren, P.; Li, M.; Luo, Z.; Song, X.; Chen, Z.; Liufu, W.; Yang, Y.; Zheng, H.; Xu, R.; Huang, Z.; et al. 2024. Infiniteworld: A unified scalable simulation framework for general visual-language robot interaction. *arXiv preprint arXiv:2412.05789*.

Rios-Martinez, J.; Spalanzani, A.; and Laugier, C. 2015. From proxemics theory to socially-aware navigation: A survey. *International Journal of Social Robotics*, 7(2): 137–153.

Sing, T. H.; and Kalingga, F. A. 2011. Gender differences in giving directions: A case study of English literature students at Binus University. *Lingua Cultura*, 5(1): 28–36.

Sun, Y.; Qiu, Y.; and Aoki, Y. 2025. DynamicVLN: Incorporating Dynamics into Vision-and-Language Navigation Scenarios. *Sensors*, 25(2): 364.

Tsoi, N.; Xiang, A.; Yu, P.; Sohn, S. S.; Schwartz, G.; Ramesh, S.; Hussein, M.; Gupta, A. W.; Kapadia, M.; and Vázquez, M. 2022. Sean 2.0: Formalizing and generating social situations for robot navigation. *IEEE Robotics and Automation Letters*, 7(4): 11047–11054.

Vasudevan, A. B.; Dai, D.; and Van Gool, L. 2021. Talk2nav: Long-range vision-and-language navigation with dual attention and spatial memory. *International Journal of Computer Vision*, 129(1): 246–266.

Vuong, A.; Nguyen, T.; Vu, M. N.; Huang, B.; Binh, H.; Vo, T.; and Nguyen, A. 2024. Habicrowd: A high performance simulator for crowd-aware visual navigation. In *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 5821–5827. IEEE.

Wang, H.; Chen, J.; Huang, W.; Ben, Q.; Wang, T.; Mi, B.; Huang, T.; Zhao, S.; Chen, Y.; Yang, S.; et al. 2024a. Grutopia: Dream general robots in a city at scale. *arXiv preprint arXiv:2407.10943*.

Wang, L.; Ma, C.; Feng, X.; Zhang, Z.; Yang, H.; Zhang, J.; Chen, Z.; Tang, J.; Chen, X.; Lin, Y.; et al. 2024b. A survey on large language model based autonomous agents. *Frontiers of Computer Science*, 18(6): 186345.

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

Wang, Z.; Zhu, Y.; Lee, G. H.; and Fan, Y. 2025. Navrag: Generating user demand instructions for embodied navigation through retrieval-augmented llm. *arXiv preprint arXiv:2502.11142*.

Wu, W.; He, H.; Wang, Y.; Duan, C.; He, J.; Liu, Z.; Li, Q.; and Zhou, B. 2024. Metaurban: A simulation platform for embodied ai in urban spaces. *arXiv e-prints*, arXiv–2407.

Xi, Z.; Chen, W.; Guo, X.; He, W.; Ding, Y.; Hong, B.; Zhang, M.; Wang, J.; Jin, S.; Zhou, E.; et al. 2025. The rise and potential of large language model based agents: A survey. *Science China Information Sciences*, 68(2): 121101.

Yang, Z.; Cai, Z.; Mei, H.; Liu, S.; Chen, Z.; Xiao, W.; Wei, Y.; Qing, Z.; Wei, C.; Dai, B.; et al. 2023. Synbody: Synthetic dataset with layered human models for 3d human perception and modeling. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 20282–20292.

Zhang, J.; Wang, K.; Xu, R.; Zhou, G.; Hong, Y.; Fang, X.; Wu, Q.; Zhang, Z.; and Wang, H. 2024. Navid: Video-based vlm plans the next step for vision-and-language navigation. *arXiv preprint arXiv:2402.15852*.

Zhou, Q.; Zhang, H.; Lin, X.; Zhang, Z.; Chen, Y.; Liu, W.; Zhang, Z.; Chen, S.; Fang, L.; Lyu, Q.; et al. 2025. Virtual community: An open world for humans, robots, and society. *arXiv preprint arXiv:2508.14893*.