LEARNING TO NAVIGATE IN OPEN URBAN ENVIRONMENTS USING A SIMPLE SIM2REAL STRATEGY

Anonymous authors
Paper under double-blind review

ABSTRACT

Autonomous navigation in open, dynamic urban environments poses unique challenges due to unstructured instructions, complex layouts, and moving obstacles. We propose Real-Nav,a unified vision-and-language navigation framework that operates seamlessly indoors and outdoors by tightly integrating semantic mapping with multimodal alignment. A simple simulation-to-reality adaptation strategy based on social-aware decision modules is employed for real-world deployment. Furthermore, in order to utilize the 3D semantic information of the space to be explored efficiently, we propose an additional pre-exploration stage in our model. We constructed a virtual environment simulator based on real photograph data, Tsinghua-roads, from Tsinghua University and completed the training on this simulator, then we evaluate Real-Nav on challenging vision-and-language navigation benchmarks and in a real-world campus setting. Our work demonstrate that building and exploiting semantic maps and employing curiosity-driven target candidate screening can significantly boost embodied navigation performance in both simulated and real-world environments.

1 INTRODUCTION

Navigating via natural language instructions in both indoor corridors and outdoor streets poses unique challenges: environments are unbounded, dynamic, and can contain distant or hard-to-recognize landmarks. Prior vision-and-language navigation (VLN) research typically treated indoor and outdoor scenarios separately, often relying on discrete graphs or pre-annotated maps that fail in complex or changing spaces.

In this work, we seek a unified approach enabling an agent to seamlessly navigate across indoor and outdoor spaces guided by language, with minimal adaptation (see Figure 1). First, a pre-exploration phase optionally leverages prior knowledge or minimal on-site scanning to initialize a coarse semantic map of the environment. This map is refined and expanded online as an RGB-D camera continuously captures new data, which is fed through a fine-tuned CLIP module to extract robust visual embeddings. Meanwhile, instruction prompts are encoded via a fine-tuned LLaMA model and merged with the visual features in a multimodal alignment step. At each timestep, the system monitors for dynamic events and, when triggered, applies a candidate scoring module to incorporate social and safety considerations. The final navigating action is selected to balance instruction fidelity and dynamic obstacle avoidance. By integrating these components, our framework ensures robust performance across indoor and outdoor environments with minimal additional domain adaptation.

2 RELATED WORK

Vision-Language Navigation. Vision-Language Navigation (VLN) has gained significant attention in recent years. The field evolves from indoor to urban settings, with ex- panded scope of tasks and datasets. Recently, the use of LLMs has introduced new solu-tions in in VLN (Zhou et al., 2024; Dorbala et al., 2024), which achieved success with indoor environment.

Sim-to-Real Transfer for Vision-Language Navigation. Anderson et al. (2018) made the first attempt at sim-to-real transfer of VLN using panoramic cameras and proposed a subgoal model based on a 2D laser scanner to identify nearby waypoints. Recent works focus on leveraging the

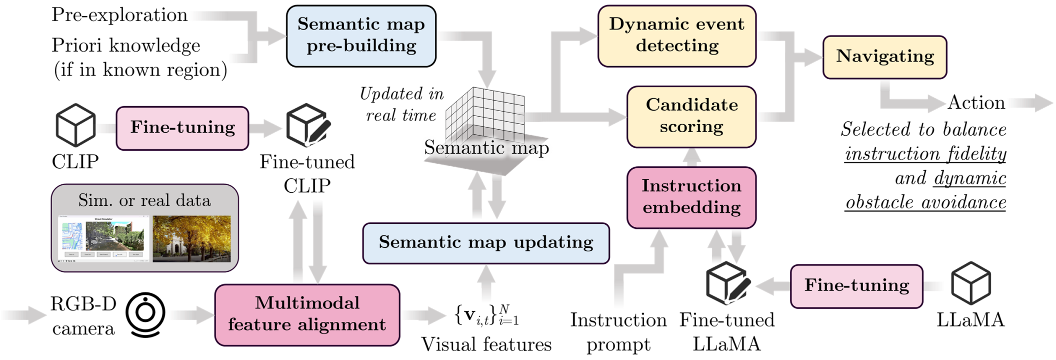

Figure 1: The work flow of Real-Nav.

generalization capabilities of LLM or multimodal models to assist in real-world VLN (Wang et al., 2025a; Qiao et al., 2025; Wang et al., 2024b; Vilera et al., 2020; Sumers et al., 2024). FLAME (Xu et al., 2025) achieved VLN by combining multiple large models or foundation models.

From indoor to outdoor. For indoor navigation, many works utilize foundational large language models based on dataset with step-by-step directions (Wang et al., 2025b)s and semantic map building (Wang et al., 2024a). Other works (Shah et al., 2022; Schumann et al., 2024) focusing on outdoor VLN, which usually use strong language understanding capabilities of LLMs for navigation based on groundlevel instructions.

3 Methods

3.1 Unified Framework for Indoor and Outdoor Navigation

Our navigation agent follows a holistic workflow that combines online mapping, vision-language understanding, and motion planning. At the core is a spatial memory updated in real-time:the agent builds a local 3D map of the environment as it moves, using onboard RGB-D camera.

Semantic Map for Waypoint Prediction. From monocular RGB-D inputs (5 cm/pixel), we generate a global occupancy map $O_t \in \mathbb{R}^{512\times512\times3}$ and semantic map $S_t \in \mathbb{R}^{512\times512\times81}$ using a pretrained UNet covering 80 object categories plus 1 background class, derived from our curated label set. A 192×192 subregion centered on the agent, concatenated with a positional embedding $P \in \mathbb{R}^{192\times192\times16}$, is refined by two UNets to yield O_t'' and S_t'' with losses

$$L_{\text{semantic}} = \text{CrossEntropy}(S_t'', S_{gt}^t), \quad L_{\text{occupancy}} = \text{CrossEntropy}(O_t'', O_{gt}^t), \tag{1}$$

These refined maps, along with PP, are input to another UNet to predict a traversability map $T_t \in \mathbb{R}^{192\times192}$ whose ground truth is optimized with

$$L_{\text{traversable}} = \text{MSE}(T_t, T_{gt}^t). \tag{2}$$

Candidate waypoints are extracted from 12 sectors (30° each) excluding occupied areas in O_t''. Additionally, a hierarchical neural radiance field (HNR) model using CLIP-ViT-B/16 [18] processes RGB-D and depth to produce a feature map. We apply HNR incrementally by caching partial 3D features at each step, and only re-render newly explored regions, ensuring sub-second updates in typical segments.

$$R \in \mathbb{R}^{8\times8\times512}, \quad R(u,v) = \sum_{n=1}^{N} \tau_n\big(1-\exp(-\sigma_n\Delta_n)\big)r_n, \quad \tau_n = \exp\Big(-\sum_{i=1}^{n-1}\sigma_i\Delta_i\Big), \tag{3}$$

which a Transformer decodes into a view representation $V \in \mathbb{R}^{1\times512}$ for integration into the VLN model.

Multimodal Feature Alignment and Candidate Scoring. Our module uses a pretrained vision-language backbone to project both text and image regions into a shared space. The natural language instruction is encoded via a 12-layer transformer to yield an embedding $e_I \in \mathbb{R}^{768}$, while candidate regions from our online 3D semantic map are encoded using a CLIP-based visual encoder to produce features $\{v_i \in \mathbb{R}^{768}\}_{i=1}^{N}$.

The alignment score is computed using cosine similarity. To encourage exploration of unknown areas, we add a curiosity bonus defined as:

$$C(i) = 1 - \exp(-\lambda u_i), \tag{4}$$

where u_i is the fraction of unknown space around candidate ii (from our rolling occupancy grid) and λ is a hyperparameter.

The final score is given by:

$$S(i) = \alpha S_{\text{VL}}(i) + \beta C(i), \tag{5}$$

with the candidate $i^* = \arg\max_i S(i)$ selected as the next navigation goal.

This process operates within the agent's egocentric 3D map—constructed by fusing monocular RGB-D data—to ensure robust feature extraction and effective alignment across domains.

3.2 Training Procedure

Our agent is trained in simulation using both imitation learning and reinforcement learning. It first learns to mimic expert trajectories—while online mapping and alignment modules continuously update the 3D semantic occupancy map—then it applies reinforcement learning to fine-tune actions based on navigation success.

To unify indoor and outdoor (Tsinghua-roads) tasks, we adopt a **multi-task objective**

$$\min_{\theta} \mathbb{E}_{\tau \sim D_{\text{indoor}} \cup D_{\text{outdoor}}} \left[\alpha \, \ell_{\text{IL}}(\theta, \tau) + \beta \, \ell_{\text{RL}}(\theta, \tau) \right], \tag{6}$$

where ℓ_{IL} is the imitation loss and ℓ_{RL} is the policy gradient loss. This formulation encourages a shared representation that is robust to both fine-grained indoor instructions and large-scale outdoor navigation.

3.3 From Simulation to Real Deployment

Our model is deployed in real urban environments using a streamlined sim-to-real transfer pipeline driven by three key components: vision-language alignment, socially-aware navigation, and world-model-based planning.

Socially-Aware Dynamic Navigation. At each timestep t, we compute the event signal as

$$e_t = \sigma(\text{MLP}(v_t - v_{t-1})), \tag{7}$$

where v_t and v_{t-1} are visual embeddings at t and $t-1$, and σ is the sigmoid function. This signal modulates cross-modal attention by forming the query $q_t = \text{Linear}_q([v_t; e_t])$ and scaling language features $k, v = \text{Linear}(w) \cdot (1 - e_t)$. When a dynamic event is detected, our system queries a pretrained vision–language model to infer a socially compliant behavior

$$B_{t+1}^{h} = (v_{t+1}^{h}, w_{t+1}^{h}) \tag{8}$$

based on the current observation and contextual prompt. We then compute a social cost for any candidate action (v_{t+1}, w_{t+1}) as

$$C_{\text{social}}^{t+1} = \lambda_l \| v_{t+1} - v_{t+1}^{h} \| + \lambda_a \| w_{t+1} - w_{t+1}^{h} \|, \tag{9}$$

where λ_l and λ_a balance the linear and angular components. This cost is integrated into our overall motion planning, steering the robot toward actions that conform to human social norms while ensuring safe and efficient navigation.

Deployment Pipeline. Deploying our model in the real world involves:

(a) Initializing the agent's map and pose (known start or via brief localization if a prior map is available). (b) Running the trained policy loop: update occupancy, align language to map and score candidates (with social cost if applicable). (c) If instruction or goal changes, the agent can re-parse and continue.

This sim-to-real transfer simplicity results from using rich, generalizable representations in sim training.

4 Experiments

4.1 Datasets and Metrics

We evaluate our work in simulation and real-world deployments. We get Tsinghua University's road data from OpenStreetMap, then apply buffering, rasterization, dilation/erosion, and image recognition to simplify the raw network. On the simplified network, we select points and fetch Baidu Maps street-view images. For indoor environments, we adopt the classic and challenging R2R benchmark.

We selected 5 paths with numerous bends and intersections in the real campus environment and placed some dynamic events. Each path underwent 10 repeated experiments. We report Success Rate (SR) and Success weighted by Path Length (SPL).

4.2 Quantitative Results

Table 1 shows our results.

Table 1: Comparison of our approach and FLAME (Xu et al., 2025) on multiple benchmarks.

Environment	Ours		FLAME	
	SR (%)	SPL (%)	SR (%)	SPL (%)
Tsinghua-roads (outdoor)	49.5	44.2	48.2	39.1
R2R (indoor)	74.5	70.4	58.4	41.8
Campus (real-environment)	45.0	42.0	28.0	22.0

Our method surpasses FLAME in all benchmarks, including a 1.3 % higher SR on Tsinghua-roads and a more substantial margin of 16.1 % on R2R, with SPL values close to SR,indicating near-optimal path planning.

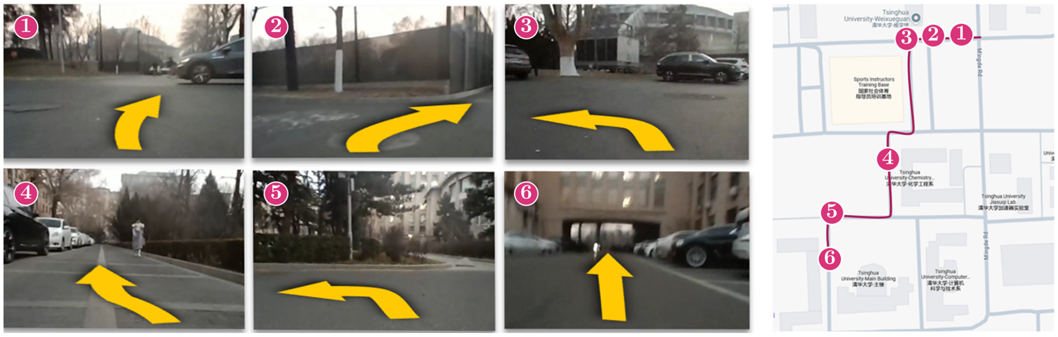
1
2
3
4
5
6
Tsinghua University-Weixueguan
清华大学·维学馆
3
2
1
Mingde Rd
Sports Instructors Training Base
国家社会体育指导员培训基地
4
Tsinghua University-Chemistry...
清华大学·化学工程系
Tsinghua University Jiasuqi Lab.
清华大学加速器实验室
5
6
Mingde Rd
Tsinghua University-Main Building
清华大学·主楼
Tsinghua University-Computer...
清华大学·计算机科学与技术系

Figure 2: Route of a campus real world trial of Real-Nav.

In the real campus deployment, Real-Nav outperforms FLAME by 17.0% SR, suggesting stronger adaptability to real-world variations.

5 Conclusion

We introduced a unified framework for vision-and-language navigation that trains in simulation and transfers seamlessly to real urban environments. Our method integrates a dynamic, social-aware module, a 3D semantic map updated in real time, and a minimal pre-exploration stage to accommodate multi-step instructions and handle real-world complexities. This work shows that lightweight sim-to-real adaptation is attainable using rich, generalizable representations derived from vision-language models and world simulation. Rather than retraining on real data, our system leverages pre-trained features to inject real-world knowledge, pointing to a future where robots are pre-adapted in simulation and readily deployable in complex, dynamic settings.

REFERENCES

Peter Anderson, Qi Wu, Damien Teney, Jake Bruce, Mark Johnson, Niko Sünderhauf, Ian Reid, Stephen Gould, and Anton Van Den Hengel. Vision-and-language navigation: Interpreting visually-grounded navigation instructions in real environments. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3674–3683, 2018.

Vishnu Sashank Dorbala, James F. Mullen, and Dinesh Manocha. Can an embodied agent find your "cat-shaped mug"? llm-based zero-shot object navigation. *IEEE Robotics and Automation Letters*, 9(5):4083–4090, 2024. doi: 10.1109/LRA.2023.3346800.

Yanyuan Qiao, Qianyi Liu, Jiajun Liu, Jing Liu, and Qi Wu. Llm as copilot for coarse-grained vision-and-language navigation. In Aleš Leonardis, Elisa Ricci, Stefan Roth, Olga Russakovsky, Torsten Sattler, and Gül Varol (eds.), *Computer Vision – ECCV 2024*, pp. 459–476, Cham, 2025. Springer Nature Switzerland.

Raphael Schumann, Wanrong Zhu, Weixi Feng, Tsu-Jui Fu, Stefan Riezler, and William Yang Wang. Velma: Verbalization embodiment of llm agents for vision and language navigation in street view, 2024. URL `https://arxiv.org/abs/2307.06082`.

Dhruv Shah, Blazej Osinski, Brian Ichter, and Sergey Levine. Lm-nav: Robotic navigation with large pre-trained models of language, vision, and action, 2022. URL `https://arxiv.org/abs/2207.04429`.

Theodore R. Sumers, Shunyu Yao, Karthik Narasimhan, and Thomas L. Griffiths. Cognitive architectures for language agents, 2024. URL `https://arxiv.org/abs/2309.02427`.

Reza Vilera, Reza Fuad Rachmadi, and Eko Mulyanto Yuniarno. Landmark segmentation and selective feature extraction in street-view image. In *2020 International Conference on Computer Engineering, Network, and Intelligent Multimedia (CENIM)*, pp. 440–444, 2020. doi: 10.1109/CENIM51130.2020.9298008.

Jiawei Wang, Teng Wang, Wenzhe Cai, Lele Xu, and Changyin Sun. Boosting efficient reinforcement learning for vision-and-language navigation with open-sourced llm. *IEEE Robotics and Automation Letters*, 10(1):612–619, 2025a. doi: 10.1109/LRA.2024.3511402.

Zehao Wang, Mingxiao Li, Minye Wu, Marie-Francine Moens, and Tinne Tuytelaars. Instruction-guided path planning with 3d semantic maps for vision-language navigation. *Neurocomputing*, 625:129457, 2025b. ISSN 0925-2312. doi: https://doi.org/10.1016/j.neucom.2025.129457. URL `https://www.sciencedirect.com/science/article/pii/S0925231225001298`.

Zihan Wang, Xiangyang Li, Jiahao Yang, Yeqi Liu, and Shuqiang Jiang. Sim-to-real transfer via 3d feature fields for vision-and-language navigation, 2024a. URL `https://arxiv.org/abs/2406.09798`.

Zihao Wang, Shaofei Cai, Guanzhou Chen, Anji Liu, Xiaojian Ma, and Yitao Liang. Describe, explain, plan and select: Interactive planning with large language models enables open-world multi-task agents, 2024b. URL `https://arxiv.org/abs/2302.01560`.

Yunzhe Xu, Yiyuan Pan, Zhe Liu, and Hesheng Wang. Flame: Learning to navigate with multimodal llm in urban environments, 2025. URL `https://arxiv.org/abs/2408.11051`.

Gengze Zhou, Yicong Hong, Zun Wang, Xin Eric Wang, and Qi Wu. Navgpt-2: Unleashing navigational reasoning capability for large vision-language models. *arXiv preprint arXiv:2407.12366*, 2024.