Scene Map-based Prompt Tuning for Navigation Instruction Generation

Supplementary Material

This document provides more details as follows:

- *Additional Quantitative Results (§A).* We provide additional quantitative results on UrbanWalk dataset and GPT-4 scoring for instruction alignment.
- *Additional Qualitative Results (§B).* Qualitative results as well as an analysis of typical cases are provided.
- *Implementation Details of* MAPINSTRUCTOR*(§C).* We introduce more implementation details of MAPINSTRUCTOR and experiment datasets. We also provide the pseudo-code of our approach for the inference procedure.
- *Discussion (§D).* We further discuss the limitations, social impact, and future work of MAPINSTRUCTOR.

A. Additional Quantitative Results

We additionally conduct experiments on the **outdoor** dataset — UrbanWalk [7] as in [5, 8, 17, 18]. UrbanWalk [7] is an outdoor navigation dataset containing 26808 image-instruction pairs simulated by CARLA [3]. We follow the experiment setting in [5, 8, 18]. Table S1 compares MAPINSTRUCTOR with several competitive models on UrbanWalk [7]. Although the dataset focuses on outdoor scenes with relatively sparse semantics, our model still achieves the best performance across all metrics. Specifically, MAPINSTRUCTOR surpasses other models by **1.4**%/**2.1**%/**2.0**%/**2.2**% in terms of SPICE/Bleu/Meteor/Rouge. This reveals the generalization capability of MAPINSTRUCTOR in outdoor scenes.

To further evaluate the quality of the generated instructions, we leverage GPT-4 [13] to score the degree of alignment between the generated instructions and the ground truth based on key landmarks and actions, using a 0-5 scale. The evaluation is conducted on REVERIE, and as shown in Table S2, our method achieves better performance, demonstrating the effectiveness of MAPINSTRUCTOR in generating highly aligned and contextually accurate navigation instructions.

Methods	UrbanWalk `test`			
	SPICE↑	Bleu-4↑	Meteor↑	Rouge↑
BT-speaker [6] [NeurIPS2018]	0.524	0.408	0.350	0.620
EDrop-speaker [15] [NAACL2019]	0.531	0.435	0.358	0.634
ASSISTER [7] [ECCV2022]	0.451	0.164	0.319	0.557
Kefa-speaker [18] [Arxiv2023]	0.566	0.450	0.378	0.655
C-INSTRUCTOR† [8] [ECCV2024]	0.645	0.534	0.461	0.781
BEVINSTRUCTOR† [5] [ECCV2024]	0.679	0.575	0.451	0.786
MAPINSTRUCTOR† **(Ours)**	**0.693**	**0.596**	**0.481**	**0.808**

Table S1. Quantitative comparison results for NIG on UrbanWalk [7] `test`. See §A for more details. †: LLM-based models.

Method	LANA+ [65]	C-INSTRUCTOR [31]	BEVINSTRUCTOR [15]	MAPINSTRUCTOR
GPT-score↑	3.21±0.4	3.48±0.3	3.57±0.4	**3.86±0.2**

Table S2. Quantitative comparison results scored by GPT-4 for NIG on REVERIE [14] `val unseen`. See §A for more details.

B. Additional Qualitative Results

Additional qualitative results are provided in Fig. S1, and Fig. S2. Fig. S1 highlights the capability of fine-grained object detection in the small, dense object environments, *e.g.*, *mirror*, *clock*, and *side table*. Fig. S2 showcases the visual results for the long-range trajectory in RxR. MAPINSTRUCTOR exhibits strong robustness in capturing crucial landmarks and temporal relationships in the long-range route, attributed to its map-based architecture design.

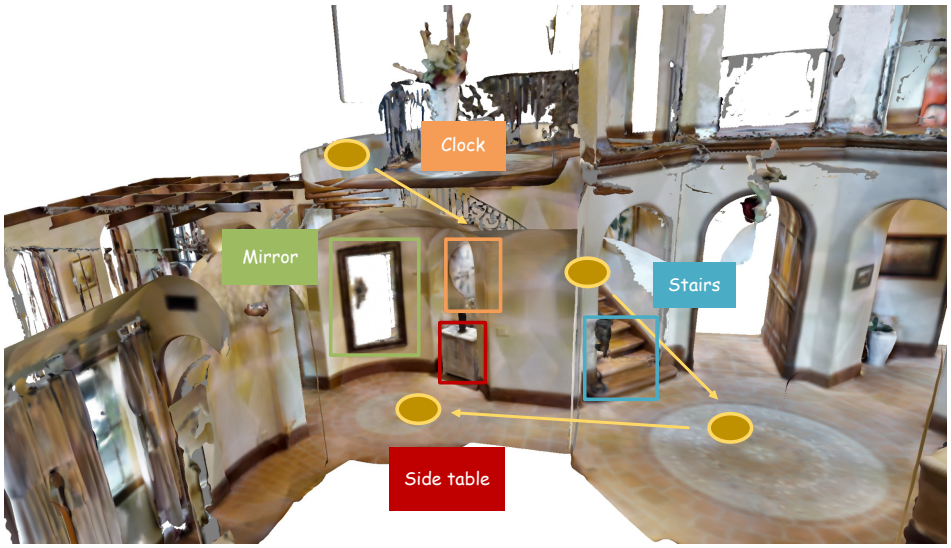
C. Implementation Details

Network Architecture. For scene representation encoding (§3.2), we utilize ViT-B/16 [4], pretrained on ImageNet, as the backbone to extract image features. Cross-View Attention is implemented using six deformable attention layers for 2D-to-3D sampling. In addition, semantic occupancy annotations are employed for multi-scale scene prediction [10, 16]. A multi-class prediction head is optimized by AdamW [12] with a learning rate of 1×10^{-4}.

Datasets. We conduct experiments on the following three datasets in the main paper.

- **R2R** [1] builds upon diverse photo-realistic house scenes. There are three splits for experiments, *i.e.*, `train` (61 scenes, 14,039 instructions), `val seen` (61 scenes, 1,021 instructions), and `val unseen` (11 scenes, 2,349 instructions). There are three human-annotated navigation instructions for each path and the average length is approximately 29 words. No overlapping scenes exist between `train` and `unseen` splits.
- **REVERIE** [14] extends Matterport3D [2] to incorporate object-level annotations. It comprises indoor scenes with 4,140 target objects and 21,702 instructions with an average length of 18 words. There are three splits for our experiment, *i.e.*, `train` (61 scenes, 10,466 instructions), `val seen` (61 scenes, 1,371 instructions), and `val unseen` (10 scenes, 3,753 instructions).
- **RxR** [9] is a multilingual dataset for Vision-Language Navigation in Matterport3D [2]. It includes longer trajectories and fine-grained visual groundings with three splits, *i.e.*, `train` (61 scenes, 11,089 instructions), `val seen` (61 scenes, 1,232 instructions), `val unseen` (10 scenes, 1,517 instructions).

We present the pseudo-code for the inference phase of MAPINSTRUCTOR in Algorithm 1.



R2R

Ground-Truth: Continue down the stairs. You'll see a big starburst tile on the floor, turn right and go into the first doorway on the right. There will be a big mirror in the room. You'll stop and wait just inside this room.

Landmark Prediction:
1. the stairs, the hallway, the doorway, the mirror
2. the stairs, the doorway, the mirror,
3. the hallway, the stairs, the first picture

Instruction Generation: walk down the stairs and turn right. walk through the doorway and stop in front of the mirror.

REVERIE

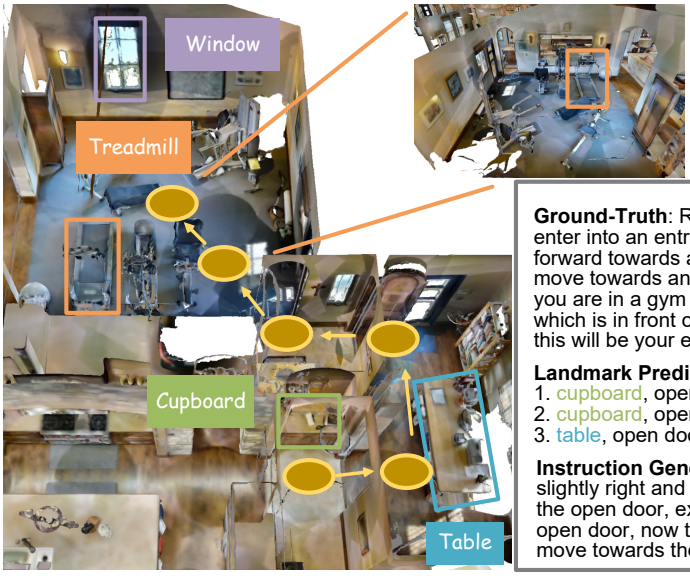
Ground-Truth: Go to the hallway with the mirror big clock and star-shaped ceiling light and dust off the side table

Goal object: side table

Landmark Prediction:
1. the hallway, the stairs, the first level
2. the stairs, the hallway, the mirror, the clock
3. the hallway, the stairs, the first level, the mirror, the clock

Instruction Generation: Go to the hallway by the stairs on the first level and clean the side table next to the mirror and clock.

Figure S1. Visual comparison results between Ground-Truth and MAPINSTRUCTOR on R2R and REVERIE. See §B for more details.



RxR

Ground-Truth: Right now you are facing a cupboard, turn slight right and enter into an entrance, which is in front of you, now turn left and move forward towards an open door, which is in front of you, now turn left and move towards an open door, now turn slight right and exit the room, now you are in a gym room, now move a little bit forward towards the window, which is in front of you, stand facing the cupboard, which is beside rack, this will be your end point.

Landmark Prediction:
1. cupboard, open door, open door, gym room, treadmill
2. cupboard, open door, open door, gym
3. table, open door, open door, gym, window, treadmill

Instruction Generation: Now you are facing towards the cupboard, turn slightly right and move forward, now turn slightly left and move towards the open door, exit the room, now turn to your left and move towards the open door, now turn to your right and move towards the gym room, move towards the treadmill.

Figure S2. Visual comparison results between Ground-Truth and MAPINSTRUCTOR for NIG on RxR. See §B for more details.

D. Discussion

Limitations. Although MAPINSTRUCTOR achieves promising performance, it remains limited by the open-vocabulary alignment between semantics and 3D representation, which leads to inaccuracies in landmark prediction. Additionally, while the 3D representation effectively captures the shapes of objects, it lacks annotations to distinguish between intra-class similar objects within the same scene. A major bottleneck for 3D-aware NIG models is the scarcity of fine-grained, diverse 3D training data, especially datasets that include detailed textual descriptions paired with 3D environments. Furthermore, NIG is a safety-critical robotic task, yet current LLM-based NIG models are prone to hallucinations. While MAPINSTRUCTOR mit-

Algorithm 1 Pseudo-code for the inference model of our approach in a PyTorch-like style

```
# F_2d: panorama perspective features
# F_3d: 3D voxel representation
# r: orientation angles
# p: perspective embedding
# a: action embedding
# v: scene representation
# v_m: map updated scene representation
# x: instruction
# s: landmark
# M: number of rounds for landmark prediction

def scene_encoder(F_2d, r):
    #= compute the perspective features (Eq.2,3) ==#
    p, a = PERSPECTIVE(F_2d, r)
    #======= compute 3D features (Eq.4,5,6) =======#
    F_3d = 3DENCODER(F_2d)
    #==== compute scene representation (Eq.7) =====#
    v = SCENE(F_3d, p, a)
    #= compute map-aggreated features (Eq.8,9,10) =#
    v_m = GNN(v)
    return v_m

def landmark_prediction(v_m):
    #====== predict landmarks (Eq.12) =======#
    s = LLM(v_m)
    return s

def instruction_generation(v_m, s):
    #=== instruction generation (Eq.1) ======#
    x = LLM(v_m, s)
    return x

def inference(F_2d, r):
     v_m = scene_encoder(F_2d, r)
     #======== recurrent refinement ==========#
    for _ in range(M):
        s = landmark_prediction (v_m)
        s_lists.append(s)
    #=== landmark semantic entropy (Eq.13) ===#
    LE(s) = ENTROPY(s_lists)
    #=========== landmark select =============#
    s_ = SELECT(s_lists)
    #======== instruction generation =========#
    x = instruction_generation(v_m, s_)

    return x
```

LLM: LLM Decoder; PERSPECTIVE: Perspective Embedding Encoder; 3DENCODER: 3D Voxel Encoder; SCENE: Scene Representation Encoder; GNN: Graph Network Encoder; ENTROPY: Landmark Semantic Entropy.

igates these hallucinations by leveraging landmark semantic entropy, it cannot eliminate the risk of generating erroneous or misleading instructions.

Social Impact. Our MAPINSTRUCTOR incorporates topological map representation into current LLM-based models via prompt features, achieving significant performance gains. This approach enhances the interactive feedback capabilities of real-world robotics, addressing the often-overlooked map connectivity in previous NIG models. It is particularly beneficial for navigation or search-and-rescue robots, especially in extreme and complex environments.

Future Work. MAPINSTRUCTOR establishes a scene representation by combining local 3D voxel representations with global topological map construction. With the advancements in modern robotics [11], embodied agents are increasingly equipped with fine-grained sensors across various modalities, such as LiDAR, IMU, and sonar. To further enhance comprehensive scene understanding, we aim to integrate additional sensors into a unified perception module, enabling richer multimodal fusion for more robust NIG. Moreover, rigorous hallucination quantification is a crucial step in real-world robotics. It directly impacts the reliability and safety of the generated instructions for agents. In the future, we will explore more effective strategies to detect, mitigate, and quantify hallucinations, ensuring higher robustness in LLM-based NIG models.

References

[1] 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 *CVPR*, 2018. 1

[2] Angel Chang, Angela Dai, Thomas Funkhouser, Maciej Halber, Matthias Niessner, Manolis Savva, Shuran Song, Andy Zeng, and Yinda Zhang. Matterport3D: Learning from RGB-D data in indoor environments. In *3DV*, 2017. 1

[3] Alexey Dosovitskiy, German Ros, Felipe Codevilla, Antonio Lopez, and Vladlen Koltun. Carla: An open urban driving simulator. In *CoRL*, 2017. 1

[4] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2020. 1

[5] Sheng Fan, Rui Liu, Wenguan Wang, and Yi Yang. Navigation instruction generation with bev perception and large language models. In *ECCV*, 2024. 1

[6] Daniel Fried, Ronghang Hu, Volkan Cirik, Anna Rohrbach, Jacob Andreas, Louis-Philippe Morency, Taylor Berg-Kirkpatrick, Kate Saenko, Dan Klein, and Trevor Darrell. Speaker-follower models for vision-and-language navigation. In *NeurIPS*, 2018. 1

[7] Zanming Huang, Zhongkai Shangguan, Jimuyang Zhang, Gilad Bar, Matthew Boyd, and Eshed Ohn-Bar. Assister: Assistive navigation via conditional instruction generation. In *ECCV*, 2022. 1

[8] Xianghao Kong, Jinyu Chen, Wenguan Wang, Hang Su, Xiaolin Hu, Yi Yang, and Si Liu. Controllable navigation instruction generation with chain of thought prompting. In *ECCV*, 2024. 1

[9] Alexander Ku, Peter Anderson, Roma Patel, Eugene Ie, and Jason Baldridge. Room-Across-Room: Multilingual vision-and-language navigation with dense spatiotemporal grounding. In *EMNLP*, 2020. 1

[10] Rui Liu, Wenguan Wang, and Yi Yang. Volumetric environment representation for vision-language navigation. In *CVPR*, 2024. 1

[11] Yang Liu, Weixing Chen, Yongjie Bai, Guanbin Li, Wen Gao, and Liang Lin. Aligning cyber space with physical world: A comprehensive survey on embodied ai. *arXiv preprint arXiv:2407.06886*, 2024. 3

[12] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *ICLR*, 2019. 1

[13] OpenAI. Gpt-4 technical report, 2023. 1

[14] Yuankai Qi, Qi Wu, Peter Anderson, Xin Wang, William Yang Wang, Chunhua Shen, and Anton van den Hengel. Reverie: Remote embodied visual referring expression in real indoor environments. In *CVPR*, 2020. 1

[15] Hao Tan, Licheng Yu, and Mohit Bansal. Learning to navigate unseen environments: Back translation with environmental dropout. In *NAACL*, 2019. 1

[16] Tai Wang, Xiaohan Mao, Chenming Zhu, Runsen Xu, Ruiyuan Lyu, Peisen Li, Xiao Chen, Wenwei Zhang, Kai Chen, Tianfan Xue, et al. Embodiedscan: A holistic multimodal 3d perception suite towards embodied ai. In *CVPR*, 2024. 1

[17] Xiaohan Wang, Wenguan Wang, Jiayi Shao, and Yi Yang. Learning to follow and generate instructions for language-capable navigation. *IEEE TPAMI*, 46(5):3334–3350, 2023. 1

[18] Haitian Zeng, Xiaohan Wang, Wenguan Wang, and Yi Yang. Kefa: A knowledge enhanced and fine-grained aligned speaker for navigation instruction generation. *arXiv preprint arXiv:2307.13368*, 2023. 1