Volumetric Environment Representation for Vision-Language Navigation Supplementary Material

Rui Liu Wenguan Wang Yi Yang*
ReLER, CCAI, Zhejiang University
https://github.com/DefaultRui/VLN-VER

This document provides more details of our approach and additional experimental results, which are organized as follows:

- Additional details (§A)
- Model details (§B)
- Discussion (§C)

A. Additional details

List of Symbols. The below table will be definitely added into our supplementary material. We will omit unnecessary subscripts for notational convenience.

Notation	Description	Index
N	Number of candidates	§3.3
$\mathcal{X}$	Volume state space	§3.2
$\mathcal{G}$	Episodic memory graph	§3.3
$\mathcal{V}$	Observed viewpoints	§3.3
$\mathcal{E}$	Navigable connections	§3.3
$\mathcal{A}$	Local action space	§3.3
$\mathcal{A}^*$	Global action space	§3.3
$\boldsymbol{E}$	Instruction embeddings	§3.2&3.3; Eq. (5)&(8)
$\boldsymbol{Q}$	3D volume query	§3.1; Eq. (1)
$\boldsymbol{F}^{2d}$	2D perspective feature	§3.1; Eq. (1)
$\boldsymbol{F}^{3d}$	3D volumetric representation	§3.1; Eq. (1)&(2)
$\boldsymbol{F}^{g}$	Height-aware group representation	§3.2; Eq. (5)
$\boldsymbol{F}^{p}$	Neighboring pillar representation	§3.3; Eq. (7)
$\boldsymbol{G}$	Node embeddings of $\mathcal{G}$	§3.3; Eq. (8)
$\boldsymbol{p}^{3d}$	Local state transition probabilities	§3.3; Eq. (4)
$\boldsymbol{p}^{2d}$	Local action probabilities	§3.3; Eq. (6)
$\boldsymbol{p}^{g}$	Global action probabilities	§3.3; Eq. (8)&(9)

† Subscript t in the paper denotes the navigation step.

Multi-resolution Labels. In coarse-to-fine VER representation extraction (§3.1), multi-resolution labels are utilized to supervise the perception network at each scale. The size of multi-resolution occupancy voxels are 0.4m, 0.2m, and 0.1m, respectively. The layout estimation and object detection are also employed at each scale. Fig. A1 shows the coarse-to-fine occupancy prediction.

Visualization. We provide more visualization results on *val unseen* of R2R [1] and REVERIE [12]. In Fig. A2, our agent recognizes the 'toilet' and 'bathtub', and then finds the first door easily. We illustrate the 3D layout estimation in Fig. A3. Given the multi-view images as input, our model can capture the 3D geometric information and reconstruct the room structure.

*Corresponding author: Yi Yang.

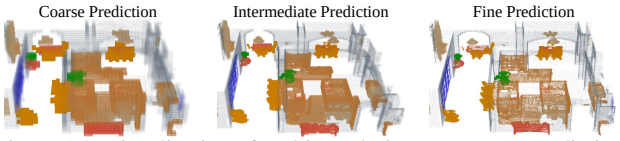

Figure A1. Visualization of multi-resolution occupancy prediction (more details in §3.1).

B. Model Details

B.1. Environment Encoder

Cross-view Attention(CVA). We propose cross-view attention for 2D-3D sampling (§3.1) and use the camera projection function $\mathcal{P}_c$ to obtain the reference points (Eq. 1), which is formulated as follows:

$$\mathrm{CVA}\big(\boldsymbol{Q}(x,y,z), \boldsymbol{F}^{2d}\big) = \mathrm{DA}\big(\boldsymbol{Q}(x,y,z), \mathcal{P}_c(\boldsymbol{p}), \boldsymbol{F}^{2d}\big), \quad \text{(B1)}$$

where $\boldsymbol{Q}(x,y,z) \in \mathbb{R}^{D_e}$ is located at (x,y,z) position of $\boldsymbol{Q} \in \mathbb{R}^{D_e \times X \times Y \times Z}$, $\boldsymbol{F}^{2d} \in \mathbb{R}^{D_i \times H \times W}$ is the image feature, and $\mathcal{P}_c$ employs the camera intrinsic and extrinsic parameters for transformation. DA is the deformable attention [14]:

$$\mathrm{DA}\big(\boldsymbol{q}, \boldsymbol{p}, \boldsymbol{F}\big) = \sum_{k=1}^{K} \boldsymbol{W}_k \sum_{s=1}^{S} \boldsymbol{A}_{ks} \boldsymbol{W}_s \boldsymbol{F}(\boldsymbol{p} + \delta \boldsymbol{p}_{ks}), \quad \text{(B2)}$$

where K is the number of attention heads, s indexes a total of S sampling points, $\boldsymbol{W}_s$ is the learning weight, $\boldsymbol{A}_{ks} \in [0,1]$ is the learnable attention weight, $\delta \boldsymbol{p}_{ks} \in \mathbb{R}^2$ is the predicted offset to the reference point $\boldsymbol{p}$, and $\boldsymbol{F}(\boldsymbol{p} + \delta \boldsymbol{p}_{ks})$ is the feature at location $\boldsymbol{p} + \delta \boldsymbol{p}_{ks}$ computed by bilinear interpolation. We sample $S = 6$ points for each query in CVA.

Multi-task Learning. We adopt ViT-B/16 [5] pretrained on ImageNet as the backbone. The size of the image features $\boldsymbol{F}^{2d}$ are $1280 \times 1024 \times 768$. We train the perception network with a detection head, a layout regression head, and a multi-class occupancy prediction head using the AdamW optimizer with a learning rate of 1×10^{-4} for 500 epoches (see §3.1).

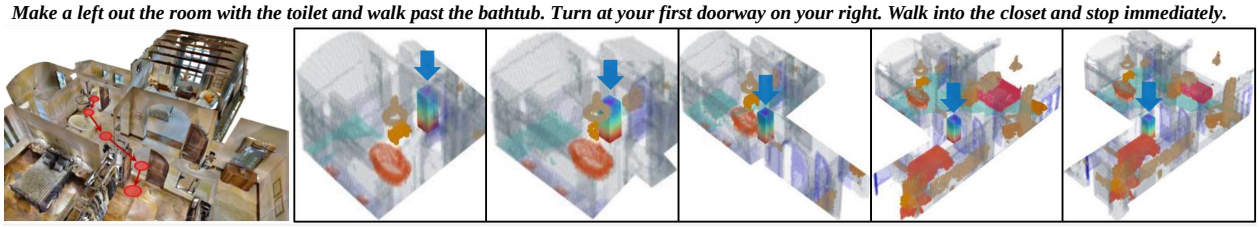

Figure A2. Visual results on *val unseen* of R2R (*i*,*ii*) and REVERIE (*iii*). During navigation, our agent recognizes the surrounding objects, captures the fine-grained details, and then performs comprehensive decision-making to finish the task successfully. Please zoom in for best view (more details in §A).

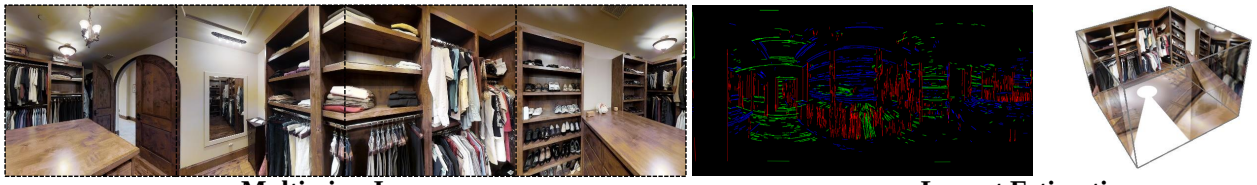

Figure A3. Visualization of 3D room layout. Different from panoramic images in previous studies [15, 16], we adopt multi-view 2D images as input. Please zoom in for best view (more details in §A).

B.2. Action Prediction

Object Prediction. For REVERIE [12], the agent needs to identify the specific objects. We first use the ViT-B/16 pre-trained on ImageNet to extract the features of N_t^o objects at t-th step $\boldsymbol{O}_t = \{\boldsymbol{o}_n | \boldsymbol{o}_n \in \mathbb{R}^{D_o}\}_{n=1}^{N_t^o}$, and add orientation features [3, 4] with heading and elevation angles ($D_o = 768$). Then these object features are concatenated with grouped VERs from different heights $\{\boldsymbol{F}_{t,z}^g \in \mathbb{R}^{D_e \times XY}\}_{z=1}^{Z}$ (§3.2). We apply multi-layer transformers (MLT) to each group as (see Eq. 5):

$$\begin{aligned} \widetilde{\boldsymbol{F}}_{t,z}^g, \widetilde{\boldsymbol{O}}_{t,z} &= \text{MLT}\big([\boldsymbol{E}; \boldsymbol{F}_{t,z}^g; \boldsymbol{O}_{t,z}]\big), \\ \widetilde{\boldsymbol{O}}_t &= \sum\nolimits_{z=1}^{Z} \widetilde{\boldsymbol{O}}_{t,z} \in \mathbb{R}^{D_o \times N_t^o}. \end{aligned} \tag{B3}$$

Then, the object prediction is formulated as:

$$\boldsymbol{p}_t^o = \text{Softmax}(\text{MLP}(\widetilde{\boldsymbol{O}}_t)) \in \mathbb{R}^{N_t^o}. \tag{B4}$$

Pretraining Objectives. For R2R [1] and R4R [7], we use Masked Language Modeling (MLM) [3, 8] and Single-step Action Prediction (SAP) [3, 6] as auxiliary tasks in the pre-training stage. For REVERIE [12], the Object Grounding (OG) [4, 10] is also used for object reasoning, and the sample ratio is MLM:SAP:OG=1:1:1. These auxiliary tasks are based on the input pair $(\boldsymbol{E}, \boldsymbol{F}_t^{\text{3d}}, \boldsymbol{G}_t)$, where $\boldsymbol{E} \in \mathbb{R}^{D_w \times L}$ are the word embeddings, $\boldsymbol{F}_t^{\text{3d}} \in \mathbb{R}^{D_e \times X \times Y \times Z}$ is the encoded VER, and $\boldsymbol{G}_t \in \mathbb{R}^{D_e \times |\mathcal{V}_t|}$ are the node embeddings of the *episodic memory* $\mathcal{G}_t$ at time step t (see §3).

Finetuning Objectives. During finetuning, we alternatively use teacher-forcing and student-forcing for action prediction [3, 6]. For REVERIE, OG is also adopted for finetuning:

$$\mathcal{L}_{action} = 0.25\mathcal{L}_{\text{tf}} + \mathcal{L}_{\text{sf}} + \mathcal{L}_{\text{OG}}. \tag{B5}$$

C. Discussion

Terms of use, Privacy, and License. Matterport3D [2], R2R [1], and REVERIE [12] are available for non-commercial research purpose. Our code is implemented on the MMDetection3D codebase. MMDetection3D (https://github.com/open-mmlab/mmdetection3d) is released under Apache 2.0 license.

Limitation. As our agent is trained and evaluated on Matterport3D Simulator, where all environments are not dynamic, deploying the algorithm directly on a real-world robot may face challenges in capturing moving objects. Therefore, additional research and development are required to ensure safe deployment in real-world scenarios. This involves incorporating flow annotations and predicting voxel velocity for foreground objects. Our work primarily addresses the interior Vision-Language Navigation task. The generalization of this approach to other navigation tasks [9, 11, 13] is not clear, and we plan to explore this in future work.

Broader Impact. We propose a powerful environment representation VER for VLN. Equipped with VER, our agent is able to perform comprehensive decision-making. On VLN benchmarks, our model demonstrates a promising improvement. In addition, we encourage more technical researching efforts devoted to environment representation learning for future research in the community.

References

[1] Peter Anderson, Qi Wu, Damien Teney, Jake Bruce, Mark Johnson, Niko Sünderhauf, Ian D. 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

[2] Angel Chang, Angela Dai, Thomas Funkhouser, Maciej Halber, Matthias Niebner, Manolis Savva, Shuran Song, Andy Zeng, and Yinda Zhang. Matterport3d: Learning from rgb-d data in indoor environments. In *3DV*, 2017. 2

[3] Shizhe Chen, Pierre-Louis Guhur, Cordelia Schmid, and Ivan Laptev. History aware multimodal transformer for vision-and-language navigation. In *NIPS*, 2021. 2

[4] Shizhe Chen, Pierre-Louis Guhur, Makarand Tapaswi, Cordelia Schmid, and Ivan Laptev. Think global, act local: Dual-scale graph transformer for vision-and-language navigation. In *CVPR*, 2022. 2

[5] 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

[6] Yicong Hong, Qi Wu, Yuankai Qi, Cristian Rodriguez-Opazo, and Stephen Gould. Vln bert: A recurrent vision-and-language bert for navigation. In *CVPR*, 2021. 2

[7] Vihan Jain, Gabriel Magalhaes, Alexander Ku, Ashish Vaswani, Eugene Ie, and Jason Baldridge. Stay on the path: Instruction fidelity in vision-and-language navigation. In *ACL*, 2019. 2

[8] Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *NAACL*, 2019. 2

[9] Jacob Krantz, Erik Wijmans, Arjun Majumdar, Dhruv Batra, and Stefan Lee. Beyond the nav-graph: Vision-and-language navigation in continuous environments. In *ECCV*, 2020. 2

[10] Xiangru Lin, Guanbin Li, and Yizhou Yu. Scene-intuitive agent for remote embodied visual grounding. In *CVPR*, 2021. 2

[11] Shubo Liu, Hongsheng Zhang, Yuankai Qi, Peng Wang, Yanning Zhang, and Qi Wu. Aerialvln: Vision-and-language navigation for uavs. In *ICCV*, 2023. 2

[12] 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, 2

[13] Karmesh Yadav, Jacob Krantz, Ram Ramrakhya, Santhosh Kumar Ramakrishnan, Jimmy Yang, Austin Wang, John Turner, Aaron Gokaslan, Vincent-Pierre Berges, Roozbeh Mootaghi, Oleksandr Maksymets, Angel X Chang, Manolis Savva, Alexander Clegg, Devendra Singh Chaplot, and Dhruv Batra. Habitat challenge 2023. https://aihabitat.org/challenge/2023/, 2023. 2

[14] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable detr: Deformable transformers for end-to-end object detection. In *ICLR*, 2020. 1

[15] Chuhang Zou, Alex Colburn, Qi Shan, and Derek Hoiem. Layoutnet: Reconstructing the 3d room layout from a single rgb image. In *CVPR*, 2018. 2

[16] Chuhang Zou, Jheng-Wei Su, Chi-Han Peng, Alex Colburn, Qi Shan, Peter Wonka, Hung-Kuo Chu, and Derek Hoiem. Manhattan room layout reconstruction from a single 360 image: A comparative study of state-of-the-art methods. *IJCV*, 129:1410–1431, 2021. 2