Supplementary for Paper: PANOGEN: Text-Conditioned Panoramic Environment Generation for Vision-and-Language Navigation

Anonymous Author(s)
Affiliation
Address
email

1 Overview

In this supplementary, we provide the following:

- Detailed description of the datasets we use in Sec. 2, and more implementation details in Sec. 3.
- More examples of the panoramic environments generated by our PANOGEN in Sec. 4.
- Limitations and broader impacts in Sec. 5, and licenses in Sec. 6.

2 Datasets

We evaluate our agent on three datasets: Room-to-Room dataset (R2R) [1], Cooperative Vision-and-Dialog Navigation dataset (CVDN) [4], and Room-for-Room dataset (R4R) [3].

R2R. Room-to-Room dataset contains detailed instructions to guide the agents navigate toward the target location step by step. The ground truth paths are the shortest path between the start location and the end location. The training set contains 61 different room environments, while the unseen validation set and test set contain 11, and 18 room environments that are unseen during training.

R4R. Room-for-Room dataset is created by concatenating the adjacent paths in the Room-to-Room dataset. In this case, the ground truth path is not the shortest path. This encourages the agent to follow the instructions to reach the target instead of exploring the environment bias and reach the target by directly navigating the shortest path.

CVDN. Cooperative Vision-and-Dialog Navigation dataset contains interactive dialogue instructions. The dialogue usually contains under-specified instructions, and the agent needs to navigate based on both the dialogue histories and the commonsense knowledge of the room. The room environments in the training set, unseen validation set, and test set follow the split in Room-to-Room dataset.

3 Implementation Details

In panoramic environment generation, we caption all the view images in the training environments in R2R dataset with BLIP-2-FlanT5-xxL. We utilize stable-diffusion-v2.1 base model to generate the single view based on caption only, and use stable-diffusion-v1.5-inpainting model to inpaint the unseen observation for the rotated views. It takes 2 days on 6 A100s to generate all the environments.

In speaker data generation, we build our speaker model based on mPLUG-base, which has 350M parameters and utilizes ViT/B-16 as the visual backbone. We train the speaker for 4 epochs on one A6000 GPU with batch size 16 for two days.

Submitted to 37th Conference on Neural Information Processing Systems (NeurIPS 2023). Do not distribute.

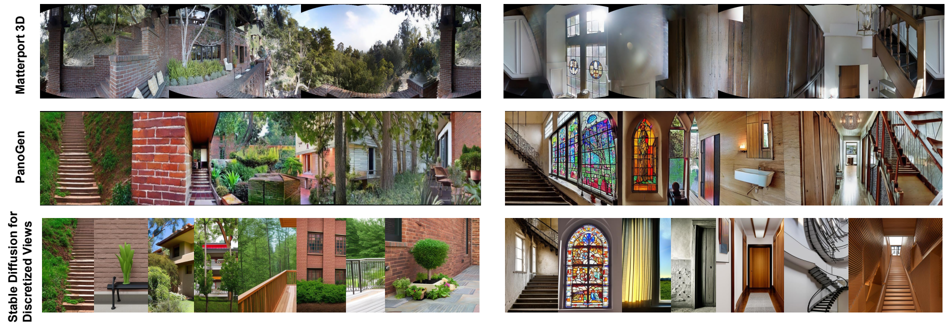

Figure 1: Qualitative analysis of the panoramic environments generated with our PANOGEN. "Matterport3D" is the original environment for VLN tasks. "Stable Diffusion for Discretized Views" is the concatenation of separately generated discretized views given text captions.

For navigation training, we adopt the agent architecture from DUET [2]. We follow the training hyperparameters in DUET. Different from DUET, we utilize CLIP-ViT/B-16 to extract the visual features. We train the model on one A6000 GPU. We pre-train the agent with batch size 64 for 150k iterations, and then fine-tune the agent with batch size 8 for 40k iterations. Both the pre-training and fine-tuning take approximately one day to finish. We report reproduced baseline performance with CLIP-ViT/B-16 features for a fair comparison. The best model is selected based on performance on validation unseen set.

4 Qualitative Example

We show more panoramic environments generated with our PANOGEN in Figure 1. We observe that directly concatenating discretized views generated separately will generate inconsistent panoramas (Row "Stable Diffusion for Discretized Views"). In comparison, our PANOGEN can generate continuous views with reasonable layout and object co-occurrence (Row "PANOGEN"). Moreover, our approach can generate panorama for both indoor and outdoor environments. Though generating the outdoor environments might not benefit agents' indoor navigation ability directly, our approach demonstrates its potential to be applied to panorama generation with different content (e.g., landscape).

5 Limitations and Broader Impacts

Vision-and-Language Navigation tasks can be used in many real-world applications, for example, a home service robot can bring things to the owner based on natural language instructions. In this paper, our proposed method generates panoramic environments for VLN training, and significantly improves navigation agents' generalization ability to unseen environments given limited human-annotated training data. Our approach reduces the efforts of re-training the agents in every new environment when adapting to real-world scenarios.

We also note that there are some limitations of our work. First, this work directly utilizes stable diffusion models trained for inpainting on "laion-aesthetics v2 5+". Though the zero-shot generation performance is good, further improvement might be observed if further trained on room images. Second, we investigate one specific task Vision-and-Language Navigation in this paper, but the proposed method can be potentially used in other embodied tasks like concept learning and grounding in panoramic environments. We will explore other useful and interesting tasks in the future.

6 Licenses

We provide the licenses of the existing assets we use in this paper in Table 1.

Table 1: A list of the licenses of the existing assets used in this paper.

Asset	**License**
Pytorch	BSD-style
Huggingface Transformers	Apache License 2.0
Torchvision	BSD 3-Clause "New" or "Revised" License
Room-to-Room	MIT
Room-for-Room	Apache License 2.0
Cooperative Vision-and-Dialog Navigation	MIT
BLIP-2	BSD 3-Clause "New" or "Revised" License
mPLUG	Apache License 2.0
DUET	N/A
Stable Diffusion	CreativeML Open RAIL-M

References

[1] 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*, pages 3674–3683, 2018.

[2] 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*, pages 16537–16547, 2022.

[3] V. Jain, G. Magalhaes, A. Ku, A. Vaswani, E. Ie, and J. Baldridge. Stay on the path: Instruction fidelity in vision-and-language navigation. *arXiv preprint arXiv:1905.12255*, 2019.

[4] J. Thomason, M. Murray, M. Cakmak, and L. Zettlemoyer. Vision-and-dialog navigation. In *Conference on Robot Learning*, pages 394–406. PMLR, 2020.