EED: Embodied Environment Description through Robotic Visual Exploration

Kohei Matsumoto
Institute of Science Tokyo
matsumoto.k.titech@gmail.com

Asako Kanezaki
Institute of Science Tokyo
kanezaki@c.titech.ac.jp

1. Introduction

While LLMs effectively provide general information and common-sense knowledge for visual navigation [2, 5, 6, 16, 17], understanding real-time events in a dynamic world remains challenging. This challenge stems, first, from the lack of an active information-gathering function, and second, from the absence of methods for searching the actively collected information. We propose a new Embodied AI task called Embodied Environment Description (EED) (Fig. 1), in which an autonomous mobile robot explores an environment and summarizes it in natural language. To properly evaluate this task, we use a crowdsourcing service to collect human-generated environment descriptions and construct a benchmark dataset. We also propose a reinforcement learning method for the robot's environment exploration behavior to perform this task, demonstrating its superior performance compared to existing visual exploration methods.

2. Embodied Environment Description Task

In an EED episode, an agent explores the environment and takes pictures. At each step t of the episode, the agent receives an egocentric RGB image $c_t \in \mathbb{R}^{256\times256\times3}$ and a corresponding depth image $d_t \in \mathbb{R}^{256\times256\times1}$ as visual observation data $o_t = (c_t, d_t)$. The agent's action space consists of three actions: FORWARD, TURN-LEFT, TURN-RIGHT. The RGB image c_t can be stored at each step. The episode length T is fixed and ends upon completing T actions. The agent then generates an environment description D_{gen} of approximately 100 words, summarizing the environment and events encountered based on the pictures. The generated environment description is evaluated using a set $\mathcal{G}$ of ground-truth descriptions (GT-Descriptions) written by humans. We propose the following three evaluation metrics suited to this specific task.

Similarity represents the similarity between the description D_{gen} generated by the agent at the end of the episode and the GT-Description $D_{\text{GT}} \in \mathcal{G}$. Both descriptions are vectorized using Sentence-BERT [13], and the average cosine similarity between these vectors is calculated as the metric.

Position-Aware Semantic Score (PAS Score) emphasizes word positioning when calculating the match between D_{gen} and D_{GT}. The PAS score is the average F1 score calculated from Precision and Recall, considering the weight of the first words in a sentence.

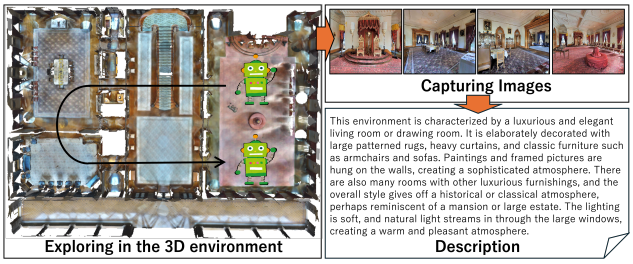
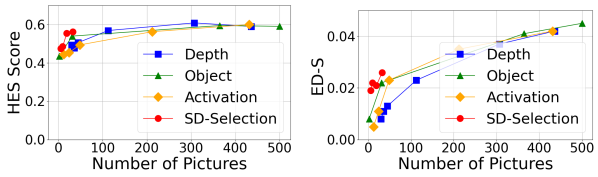


Figure 1. Overview of Embodied Environment Description (EED).

Human-Enhanced Similarity Score (HES Score) replicates human evaluation of environment descriptions by fine-tuning an LLM. We fine-tune Sentence-BERT [13] using a dataset of human evaluations of environment descriptions we collected. The output values are normalized to $[0.0, 1.0]$. It is experimentally confirmed that the correlation between the HES score and human evaluation is 0.840, which is significantly higher than the BLEU score of 0.548.

2.1. Data Collection through Crowdsourcing

We used the crowdsourcing platform Amazon Mechanical Turk (AMT) to create a set $\mathcal{G}$ of five GT-Descriptions for each of 90 scenes in Matterport 3D [4]. Subsequently, we collected human evaluation data for environment descriptions. Using a custom browser-based tool we developed, workers watched videos that comprehensively explored Matterport 3D scenes in Habitat-Sim [14], wrote environment descriptions, and evaluated those descriptions. These data were collected through the platform. We manually filtered poor-quality data. The set of GT-Descriptions $\mathcal{G}$ for each scene consists of five descriptions: the four descriptions collected as described above, along with one additional description created by us. For the collection of human evaluation data, we selected 20 descriptions from each of 11 scenes for evaluation. Each worker rated a set of five displayed descriptions on a 5-point Likert scale from 1 to 5.



(a) HES score (b) ED-S [3]

Figure 2. Comparison of SD-Selection and Speaker Policy [3].

Table 1. Experimental results for the EED Task.

Method	HES score	PAS score	Similarity	BLEU	ROUGE-1	METEOR	Explored Area
Random	0.740	0.482	0.584	0.021	0.320	0.208	0.152
Novelty	0.760	0.503	0.592	0.022	0.323	0.212	0.169
Coverage	0.756	0.495	**0.596**	0.021	0.326	0.210	**0.308**
Smooth-Coverage [12]	0.752	0.489	0.589	0.021	0.325	0.209	0.202
Curiosity	0.746	0.469	0.588	0.022	0.329	0.211	0.153
Reconstruction	0.714	0.501	0.583	**0.022**	**0.339**	0.211	0.160
EnvDescriber (Ours)	**0.767**	**0.512**	0.595	0.021	0.330	**0.212**	0.206
Human Baseline	0.781	0.532	0.630	0.022	0.333	0.222	-

The evaluation score for each environment description was calculated as the average of scores collected from 10 different workers. Krippendorff's α [7], which measures the degree of agreement between annotators, was 0.320, indicating that our collected data is highly reliable [10].

2.2. Method

Picture Selection. We define the value v_p for each picture $p \in P$, and based on this value and the similarity between pairs of pictures, the selection is made to create the set of pictures P_{select} required for generating the environmental description. First, the value of a picture is defined as $v_p = N_p^{\text{sal}} \times N_p^{\text{cat}}$. In this equation, N_p^{sal} is the number of pixels in the predicted saliency map $\hat{s}_p$ of picture p (generated by TranSalNet [9]) with a value of 0.5 or higher, and N_p^{cat} is the number of object categories present in p. Then, using v_p and the similarity $\text{Sim}(p_i, p_j)$ between p_i and p_j computed by CLIP [11], we construct the set P_{select}. From now on, we refer to this method of selecting pictures as Saliency-Driven Selection (SD-Selection).

Environmental Description Generation. The agent generates the environmental description D_{gen} by inputting the set of pictures selected P_{select} into a VLM. We use LLaVA [8] as the VLM. First, each picture $p_i \in P_{\text{select}}$ is input to the VLM to generate the description D_{p_i} for picture p_i. Next, the pictures in P_{select} are arranged in a two-column layout to form a single image, which is then input to the VLM, resulting in the output of the environmental description D_{gen}.

EnvDescriber: EnvDescriber is the exploration method proposed in this paper and serves as the baseline for the EED Task. The agent is trained using a reward structure that combines four reward components, defined as $r_t = r_{\text{HES}} + r_{\text{area}} + r_{\text{sub-goal}} + r_{\text{time-penalty}}$. In this equation, r_{HES} is provided only at the end of an episode and corresponds to the HES score calculated between the generated environment description and the GT set $\mathcal{G}$. r_{area} is a reward for spatial exploration. $r_{\text{sub-goal}}$ is computed based on whether the agent observes objects mentioned in $\mathcal{G}$ at time t. Finally, $r_{\text{time-penalty}}$ is a negative slack reward.

3. Results and Discussion

Picture Selection Method Comparison. Figure 2 shows a comparison between our proposed SD-Selection method and the threshold-based picture selection method called Speaker Policy, proposed by Bigazzi et al. [3], which is the most closely related prior work. The comparison uses the HES score and ED-S [3]. From the figure, we observe a general trend that the evaluation scores increase as the number of selected pictures grows. Notably, SD-Selection is plotted in the upper-left region for both metrics, indicating that it performs better than Speaker Policy with fewer selected pictures. Although Speaker Policies can exceed SD-Selection by increasing the number of pictures, it is necessary to select more than 100 pictures in some cases, making it difficult to input into a time-consuming VLM.

Performance Evaluation of the EED Task. Our dataset consists of Matterport3D scenes, with a standard train/val/test split of 61/11/18 scenes. Following the recommendations of Anderson *et al*. [1], there is no overlap between the train, validation, and test scenes. The agent's initial position is randomly sampled. All agents train with proximal policy optimization (PPO) [15]. In addition to the evaluation metrics described in Section 2, we evaluate *Explored Area*, which is the rate of the area has already been explored by the agent as a supplementary evaluation measure. This experiment compares EnvDescriber with a randomly acting agent and agents trained with the five exploration reward structures introduced by Ramakrishnan *et al*. [12]. Additionally, we compare the performance of *Human Baseline*, where five humans performed the exploration and picture selection, using the average performance as a reference. Table 1 shows the performance of each method in the validation scene episodes. This table shows that *Random* exhibited the lowest performance across all metrics, confirming that appropriate exploration behavior and picture selection are crucial for achieving good performance in the EED task. Additionally, from the result of the highest *Explored Area* in *Coverage*, it is suggested that simply exploring the environment broadly is insufficient, and how the agent moves (and collects information) is essential for performance in the EED task. Furthermore, EnvDescriber achieved the highest values for both the HES score and PAS score, critical indicators in the EED task. This indicates that focusing on visiting distinctive locations within the environment and actively collecting information can contribute to improved performance in the EED task. You can access our dataset at https://github.com/ak-lab-titech/EED.

References

[1] Peter Anderson, Angel Chang, Devendra Singh Chaplot, Alexey Dosovitskiy, Saurabh Gupta, Vladlen Koltun, Jana Kosecka, Jitendra Malik, Roozbeh Mottaghi, Manolis Savva, et al. On evaluation of embodied navigation agents. *arXiv preprint arXiv:1807.06757*, 2018. 2

[2] 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 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3674–3683, 2018. 1

[3] Roberto Bigazzi, Marcella Cornia, Silvia Cascianelli, Lorenzo Baraldi, and Rita Cucchiara. Embodied agents for efficient exploration and smart scene description. In *Proceedings of IEEE International Conference on Robotics and Automation (ICRA)*, pages 6057–6064. IEEE, 2023. 2

[4] 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 *Proceedings of Conference on 3D Vision (3DV)*, pages 667–676. IEEE Computer Society, 2017. 1

[5] 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. *Proceedings of Advances in Neural Information Processing Systems (NeurIPS)*, 31, 2018. 1

[6] Chenguang Huang, Oier Mees, Andy Zeng, and Wolfram Burgard. Visual language maps for robot navigation. In *Proceedings of IEEE International Conference on Robotics and Automation (ICRA)*, pages 10608–10615. IEEE, 2023. 1

[7] K Krippendorff. Content analysis: an introduction to its methodology. *The Sage commtext series (5) Show all parts in this series*, 1980. 2

[8] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Proceedings of Advances in Neural Information Processing Systems (NeurIPS)*, 36, 2024. 2

[9] Jianxun Lou, Hanhe Lin, David Marshall, Dietmar Saupe, and Hantao Liu. Transalnet: Towards perceptually relevant visual saliency prediction. *Neurocomputing*, 494:455–467, 2022. 2

[10] Mayu Otani, Riku Togashi, Yu Sawai, Ryosuke Ishigami, Yuta Nakashima, Esa Rahtu, Janne Heikkilä, and Shin ' ichi Satoh. Toward verifiable and reproducible human evaluation for text-to-image generation. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14277–14286, 2023. 2

[11] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *Proceedings of International Conference on Machine Learning (ICML)*, pages 8748–8763. PMLR, 2021. 2

[12] Santhosh K Ramakrishnan, Dinesh Jayaraman, and Kristen Grauman. An exploration of embodied visual exploration. *International Journal of Computer Vision*, 129(5): 1616–1649, 2021. 2

[13] N Reimers. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*, 2019. 1

[14] Manolis Savva, Abhishek Kadian, Oleksandr Maksymets, Yili Zhao, Erik Wijmans, Bhavana Jain, Julian Straub, Jia Liu, Vladlen Koltun, Jitendra Malik, et al. Habitat: A platform for embodied ai research. In *Proceedings of International Conference on Computer Vision (ICCV)*, pages 9339–9347, 2019. 1

[15] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017. 2

[16] Dhruv Shah, Błażej Osiński, Sergey Levine, et al. Lm-nav: Robotic navigation with large pre-trained models of language, vision, and action. In *Proceedings of Conference on Robot Learning (CoRL)*, pages 492–504. PMLR, 2023. 1

[17] Mohit Shridhar, Jesse Thomason, Daniel Gordon, Yonatan Bisk, Winson Han, Roozbeh Mottaghi, Luke Zettlemoyer, and Dieter Fox. Alfred: A benchmark for interpreting grounded instructions for everyday tasks. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10740–10749, 2020. 1