KERM: Knowledge Enhanced Reasoning for Vision-and-Language Navigation — Supplementary Material —

Xiangyang Li[1,2], Zihan Wang[1,2], Jiahao Yang[1,2], Yaowei Wang[3], Shuqiang Jiang[1,2,3]
[1]Key Lab of Intelligent Information Processing Laboratory of the Chinese Academy of Sciences (CAS), Institute of Computing Technology, Beijing, 100190, China
[2]University of Chinese Academy of Sciences, Beijing, 100049, China
[3]Peng Cheng Laboratory, Shenzhen, 518055, China
lixiangyang@ict.ac.cn, {zihan.wang, jiahao.yang}@vipl.ict.ac.cn, wangyw@pcl.ac.cn, sqjiang@ict.ac.cn
https://github.com/XiangyangLi20/KERM

Section 1 provides additional details for the pretraining of our overall framework. The experimental results for the improvement based on the fine-scale encoder and the coarse-scale encoder are illustrated in Section 2. Section 3 presents more qualitative examples.

1. Pretraining Details

To pretrain our proposed KERM, we utilize four auxiliary tasks. Besides the behavior cloning tasks, *i.e.,* single-step action prediction (SAP) and object grounding (OG), the masked language modeling (MLM) and masked region classification (MRC) are utilized. In the following, these two tasks are described.

- Masked language modeling (MLM). MLM aims to learn language representations by masking parts of the text and predicting them with the contextual information. The inputs of this task are pairs of language instruction L and the corresponding demonstration path P. As our method utilizes the dual-scale graph transformer [2] for action prediction, we also average the embeddings of the fine-scale and coarse-scale encoders and then a network with two fully-connected layers is used to predict the target word. Similar to previous approaches [4, 6], we randomly mask out the instruction words with a probability of 15%. This task is optimized by minimizing the negative log-likelihood of the original words:

$$L_{MLM} = -log(w_i|L_m, P)$$

  where L_m is the masked instruction and w_i is the label of the masked word.

- Masked region classification (MRC). MRC requires the model to predict the semantic labels of masked view images according to the instruction, unmasked view images, and the corresponding features in the topological map. With the same settings in DUET [2], we randomly mask out view images and objects in the last observation of the corresponding demonstration path P with a probability of 15% in the fine-scale encoder. The visual features for the masked images or objects are set to zero, while their position embeddings are preserved. The target semantic labels for view images are predicted by an image classification model [3] pretrained on ImageNet, and the labels for the objects are obtained by an object detector [1] pretrained on the Visual Genome dataset [5]. Similar to [2], we use a two-layer fully-connected network to predict the semantic labels of masked visual tokens, and the KL divergence between the predicted and target probability distribution of each mask token is minimized.

2. Improvements on Different Scales

We also evaluate our proposed method on the settings with only the fine-scale encoder and the coarse-scale encoder separately, as illustrated in Table 1. When only using the fine-scale encoder to predict the action, our KERM-fine significantly outperforms DUET-fine. For example, the Success Rate (SR) is improved from 28.86% to 30.80%. The trend on the coarse-scale encoder is also the same. With both the fine-scale and the coarse-scale encoders, our KERM improves the SR by 3.3%. The results demonstrate the effectiveness of our method.

Furthermore, we investigate the effect of the strategy that the agent selects the most likely viewpoint in the navigation history if the timesteps go above a certain threshold. For the fair comparison with previous approaches, we apply the same settings as [2]. Specifically, we set the maximum ac-

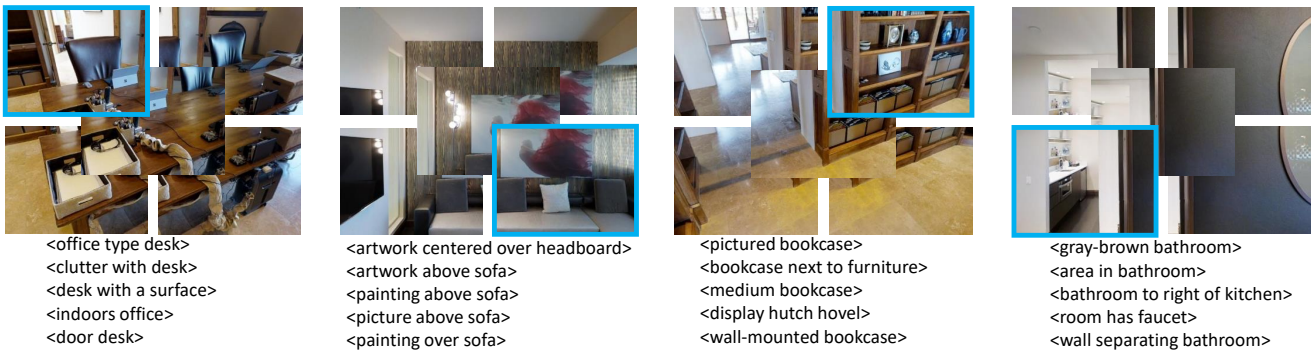

Figure 1. Examples of the retrieved facts for view images. Each view image is cropped into five sub-regions. We show the top-5 facts for the sub-regions in the blue box.

Table 1. The results of different scales and dual-scale fusion on the val unseen split of the REVERIE dataset.

	OSR↑	SR↑	SPL↑	RGS↑	RGSPL↑
DUET-fine	30.96	28.86	23.57	20.39	16.64
KERM-fine	34.40	30.80	24.83	21.68	17.56
DUET-coarse	46.44	36.52	25.98	-	-
KERM-coarse	46.38	37.38	26.32	-	-
DUET	51.07	46.98	33.73	32.15	23.03
KERM	55.21	50.44	35.38	34.51	24.45

Table 2. Statistics of episodes with the maximum action steps.

	Val Seen	Val Unseen	Test
REVERIE	31/1423 (2.18%)	560/3521 (15.90%)	898/6292 (14.27%)
R2R	10/1021 (0.98%)	51/2349 (2.17%)	132/4173 (3.16%)
SOON	-	888/2261 (39.27%)	1423/3080 (46.20%)

tion steps as 15 for REVERIE and R2R, and 20 for SOON. Table 2 illustrates the proportion of the episodes terminated with the maximum action steps. For example, on the val unseen split of REVERIE, "560/3521 (15.90%)" represents that this split has 3521 episodes, and 560 of them are terminated by the criterion of the termination policy (TP), making up a proportion of 15.90%. Moreover, Table 3 shows the results of our KERM and the policy that the agent stops at the last visited location (*i.e.*, KERM w/o TP). The results illustrate that the employed TP has small influence on REVERIE and R2R, while has great influence on SOON. This is because that the average hop of the trajectories on SOON is longer with more complex language instructions.

3. More Qualitative Results

Figure 1 illustrates the retrieved facts for view images. The retrieved facts provide crucial information (*e.g.,* attributes and relationships between objects) which are complementary to visual features. Figure 2 demonstrates the

Table 3. Influence of the termination policy on the val unseen split.

		OSR	SR	SPL	RGSPL
REVERIE	KERM	**55.21**	**50.44**	**35.38**	**24.45**
	KERM w/o TP	55.21	49.96	35.29	24.25
R2R	KERM	**80.42**	**71.95**	**60.91**	-
	KERM w/o TP	80.42	71.90	60.77	-
SOON	KERM	**51.62**	**38.05**	**23.16**	**4.04**
	KERM w/o TP	51.62	35.23	21.53	3.52

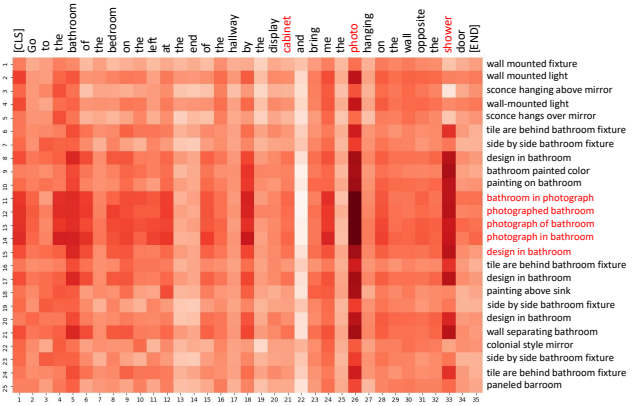

Figure 2. Illustration of the weights for the 25 facts during fact purification. Best viewed in color.

weights for each fact corresponding to each word in the instruction during fact purification. It is illustrated that our model can automatically select the relevant facts to make better action prediction.

References

[1] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. Bottom-up and top-down attention for image captioning and visual question answering. In *CVPR*, pages 6077–6086, 2018. 1

[2] 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*, pages 16537–16547, 2022. 1, 2

[3] 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

[4] Weituo Hao, Chunyuan Li, Xiujun Li, Lawrence Carin, and Jianfeng Gao. Towards learning a generic agent for vision-and-language navigation via pre-training. In *CVPR*, pages 13137–13146, 2020. 1

[5] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *IJCV*, 123(1):32–73, 2017. 1

[6] Yanyuan Qiao, Yuankai Qi, Yicong Hong, Zheng Yu, Peng Wang, and Qi Wu. HOP: History-and-order aware pre-training for vision-and-language navigation. In *CVPR*, pages 15418–15427, 2022. 1