Scene-Intuitive Agent for Remote Embodied Visual Grounding (Supplementary Material)

Xiangru Lin[1] Guanbin Li[2*] Yizhou Yu[1,3]

[1]The University of Hong Kong [2]Sun Yat-sen University [3]Deepwise AI Lab

xrlin2@cs.hku.hk, liguanbin@mail.sysu.edu.cn, yizhouy@acm.org

1. More Related Work

Behavioral Research on Human Navigation. The behavioural research of navigation of human beings has a long history and is still under active research [14, 6, 16, 15, 1, 2, 4]. Yet, it is not well understood how we human carry out the learning process of navigation in our brain to allow us to navigate in a familiar or unfamiliar environment. However, according to [14, 16, 15, 1], we humans use a range of different cognitive processes when we navigate. For example, we identify representative landmark cues, memorize our goal location, and identify the shortest route to that goal location. A significant number of research have supported such dissociable cognitive aspects. The human intuitions for remote embodied navigation we referred to in this paper is a set of commonsense rules and heuristics that come from observations of humans life experiences, which shares a similar motivation mentioned in [5]. Our work has also proved that drawing on such observations in high-level VLN is a promising direction.

2. Implementation Details

In this section, we introduce the implementation details of the pre-training stage and the action decoding stage. In pre-training stage, we first present the sampled datasets information of the Scene Grounding task and the Object Grounding task. Second, we introduce the ViLBERT model used in the pre-training stage. Third, we illustrate the action decoder architecture and the training parameters in detail.

2.1. Pre-training Stage Details

Scene Grounding Task. The Scene Grounding training dataset consists of 10312 samples, each containing an instruction and four viewpoints out of which one is positive. The sampling strategy is illustrated in the main paper. We evaluate the effectiveness of this task by asking the model trained to identify the true target viewpoint given the ground-truth path. We report the accuracy on the Val Seen (1423 paths) and Val UnSeen (3521 paths) REVERIE.

Object Grounding Task. The image based grounding dataset contains 67432 training samples and the viewpoint based object grounding dataset contains 4356 training samples. The sampling strategy is presented in the main paper. Similar to [12], we evaluate the performance of this model on the ground-truth target viewpoint and report the object grounding accuracy.

Model Details. The ViLBERT model used in Scene Grounding task and Object Grounding task consists of a language stream, a vision stream and a cross modal alignment layers block. The language stream utilizes a $BERT_{BASE}$ architecture [3], which has 12-layer of transformer blocks and each block having a hidden state size of 768 and 12 attention heads. The vision stream and the cross modal alignment block use 6-layer transformer blocks and each having a hidden state size of 1024 and 8 attention heads respectively. Following [8, 9], the language stream is initialized with BERT weights pre-trained on the BookCorpus [18] and English Wikipedia datasets. Then, the ViLBERT model is pre-trained on the Conceptual Captions dataset [11] as well as the 12 tasks specified in [9]. Finally, it is fine-tuned on our Scene Grounding task and Object Grounding task respectively. In the Scene Grounding task, the Scene Grounding model is trained with the Adam optimizer with a learning rate of $4e-5$ and a batch size of 32 for 10 epochs. In the Object Grounding task, the Object Grounding model is first trained on the image based Object Grounding dataset with the Adam optimizer with a learning rate of $4e-5$ and a batch size of 128 for 20 epochs. Then, it is further fine-tuned on the viewpoint based Object Grounding dataset with the Adam optimizer with a learning rate of $1e-5$ and a batch size of 128 for 10 epochs. We use a linear decay learning rate schedule with warm up to train the aforemen-

*Corresponding author is Guanbin Li. This work was supported in part by the National Key Research and Development Program of China (No.2020YFC2003902), in part by the Guangdong Basic and Applied Basic Research Foundation under Grant No.2020B1515020048, in part by the National Natural Science Foundation of China under Grant No.61976250 and No.U1811463. This work was also sponsored by CCF-Tencent Open Research Fund.

tioned models. All models are trained on NVIDIA Geforce 2080Ti GPUs with 11GB memory using Pytorch [10].

2.2. Action Decoding Stage Details

The $ViLEncoder$ is composed of a ViLBERT model pre-trained on the Scene Grounding task and a Bi-directional LSTM layer. The D_h in the BiLSTM is set to 512. The $ViLPointer$ is ViLBERT model pre-trained on the Object Grounding task. The N_{mem} and N_{state} used in the memory blocks are set to 3 according to our ablation study in the ablation study. We follow the same RL setting as [13] that sets the discounted factor to 0.9 and adopts reward shaping [17]. We train the agent with the Adam Optimizer [7] with a learning rate of $1e-4$, weight decay of $5e-4$, batch size of 64 and the maximum decoding action length of 40. We clip the global gradient norm at 40. We train the agent for 13000 iterations and report the final performance. All experiments have been conducted on NVIDIA Geforce 2080Ti GPUs with 11GB memory using Pytorch [10].

3. Evaluation Metrics Details

In this section, we illustrate the details of the evaluation metrics. Following [12], we evaluate the performance of the model based on REVERIE Success Rate (RGS) and REVERIE Success Rate weighted by Path Length (RG SPL). Besides, we report the performance of our method on the following metrics in the REVERIE dataset. It is worth noting that the target object is only observable within 3 meters of the target viewpoint.

- **Navigation Success Rate** is the percentage of the target object observable at the agent's final location.
- **Navigation Oracle Success Rate** measures the percentage of the target object that can be observed at one of the agent's passed viewpoints.
- **Navigation Success Rate weighted by Path Length (SPL)** is the navigation success rate weighted by the trajectory length.
- **Navigation Length** is the trajectory length in meters.
- **REVERIE Success Rate (RGS)** is calculated as the percentage of the output bounding box that has an IoU ≥ 0.5 with the ground truth box.
- **REVERIE Success Rate weighted by Path Length (RG SPL)** is REVERIE success rate weighted by the trajectory length.

4. Qualitative Examples

In this section, we show a number of qualitative examples of how our proposed agent performs in both Val Seen environment (from Fig. 1 to Fig. 4) and Val Unseen environment (from Fig. 5 to Fig. 8). Besides, we also visualize five representative failed cases illustrating the typical mistakes our agent make to better understand how our agent works.

References

[1] Elizabeth R. Chrastil. Neural evidence supports a novel framework for spatial navigation. *Psychonomic Bulletin & Review*, 20:208–227, 2013. 1

[2] A. Coutrot, R. Silva, E. Manley, Will de Cothi, and H. Spiers. Global determinants of navigation ability. *Current Biology*, 28:2861–2866.e4, 2018. 1

[3] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018. 1

[4] S. Gopal, R. Klatzky, and T. Smith. Navigator: A psychologically based model of environmental learning through navigation. *Journal of Environmental Psychology*, 9:309–331, 1989. 1

[5] Saurabh Gupta, Varun Tolani, J. Davidson, S. Levine, R. Sukthankar, and J. Malik. Cognitive mapping and planning for visual navigation. *International Journal of Computer Vision*, 128:1311–1330, 2019. 1

[6] Simon Jetzschke, M. Ernst, J. Fröhlich, and N. Boeddeker. Finding home: Landmark ambiguity in human navigation. *Frontiers in Behavioral Neuroscience*, 11, 2017. 1

[7] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2015. 2

[8] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In *Advances in Neural Information Processing Systems*, pages 13–23, 2019. 1

[9] Jiasen Lu, Vedanuj Goswami, Marcus Rohrbach, Devi Parikh, and Stefan Lee. 12-in-1: Multi-task vision and language representation learning. In *The IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 1

[10] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems 32*, pages 8024–8035. 2019. 2

[11] Sharma Piyush, Ding Nan, Goodman Sebastian, and Soricut Radu. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, 2018. 1

[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 *Proceedings of the*

IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2020. 1, 2

[13] Hao Tan, Licheng Yu, and Mohit Bansal. Learning to navigate unseen environments:back translation with environmental dropout. In *Proceedings of The North American Chapter of the Association for Computational Linguistics (NAACL)*, 2019. 2

[14] I. V. D. van der Ham, M. H. G. Claessen, A. Evers, and Milan N. A. van der Kuil. Large-scale assessment of human navigation ability across the lifespan. *Scientific Reports*, 10, 2020. 1

[15] J. Wiener, Simon J. Büchner, and C. Hölscher. Taxonomy of human wayfinding tasks: A knowledge-based approach. *Spatial Cognition & Computation*, 9:152 – 165, 2009. 1

[16] T. Wolbers and M. Hegarty. What determines our navigational abilities? *Trends in Cognitive Sciences*, 14:138–146, 2010. 1

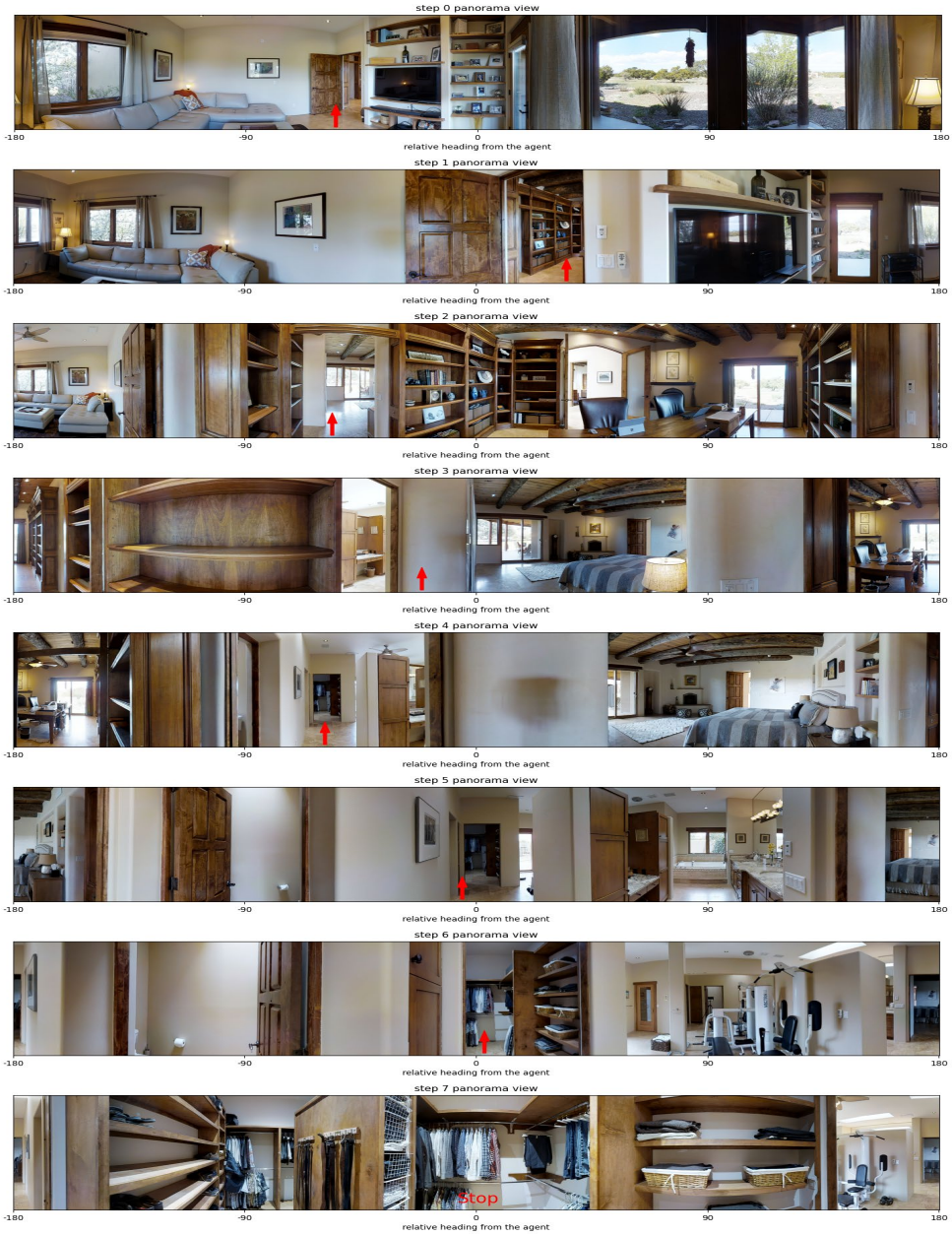
[17] Yi Wu, Yuxin Wu, Georgia Gkioxari, and Yuandong Tian. Building generalizable agents with a realistic and rich 3d environment. *arXiv preprint arXiv:1801.02209*, 2018. 2

[18] Yukun Zhu, Ryan Kiros, Rich Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, December 2015. 1

Instruction

Go to the closet and give me the pair of blue shoes that are on the third shelf from the bottom

Navigation steps of the scene-intuitive agent. The **red** arrow shows the action chosen by the agent. The **Stop** sign means the agent stops at corresponding viewpoint.



The **red** bounding box denotes the localized output object and the **green** bounding boxes are candidate objects.

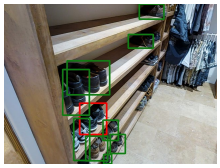


Figure 1. The successful navigation and localization qualitative example result on Val Seen dataset.

Instruction

Go to your right when entering the front door and enter the office on your left with a big desk in it and no doors to the room and stand at the black desk in the middle of the office

Navigation steps of the scene-intuitive agent. The **red** arrow shows the action chosen by the agent. The **Stop** sign means the agent stops at corresponding viewpoint.

The **red** bounding box denotes the localized output object and the **green** bounding boxes are candidate objects.

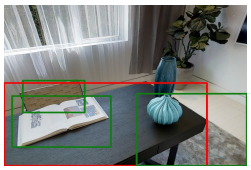
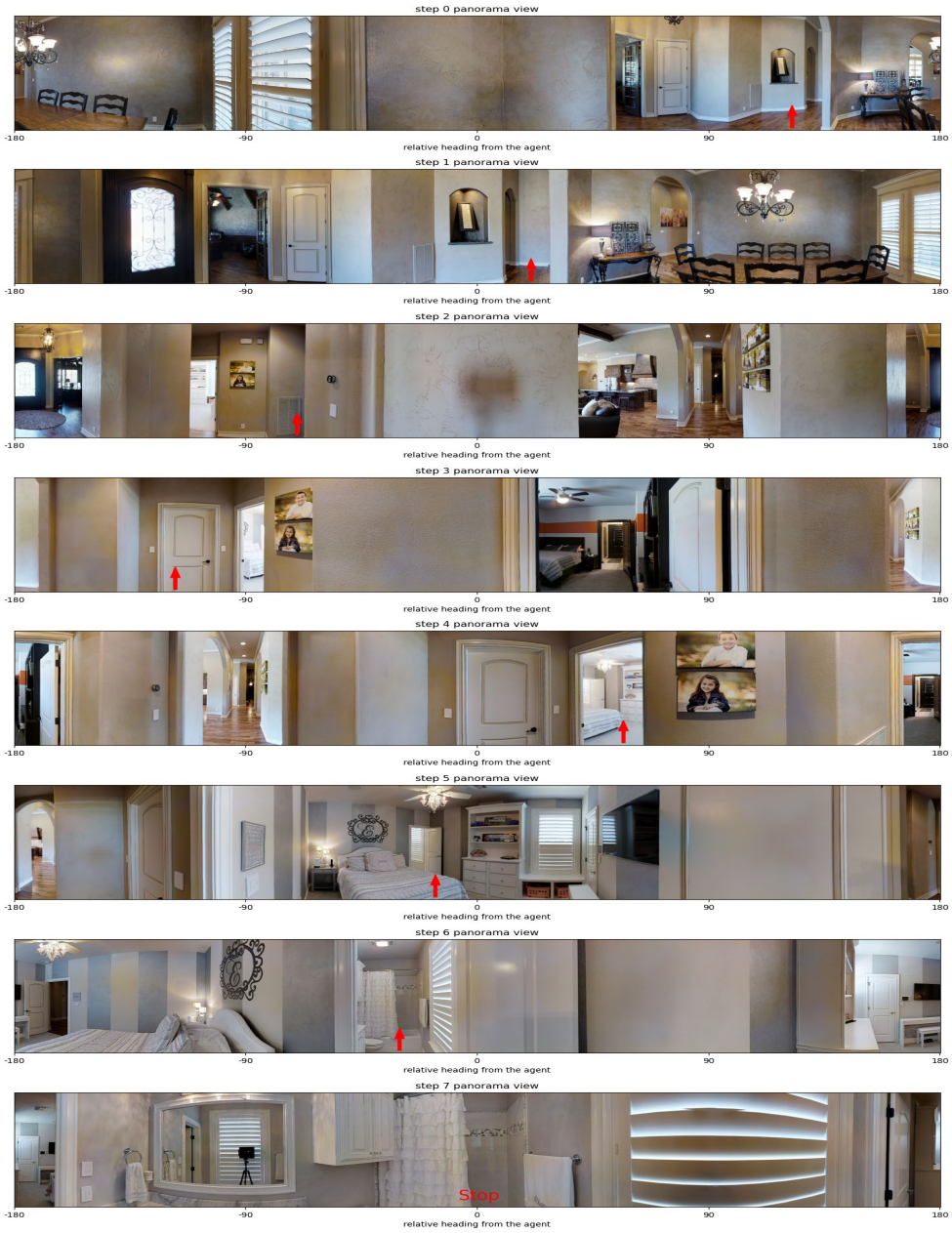


Figure 2. The successful navigation and localization qualitative example result on Val Seen dataset.

Instruction

Go to the bathroom with the white ruffle shower curtain and Elliana monogrammed on the towel off the bedroom with the big metal E behind the bed and check if there is any toilet paper on the holder next to the toilet

Navigation steps of the scene-intuitive agent. The **red** arrow shows the action chosen by the agent. The **Stop** sign means the agent stops at corresponding viewpoint.



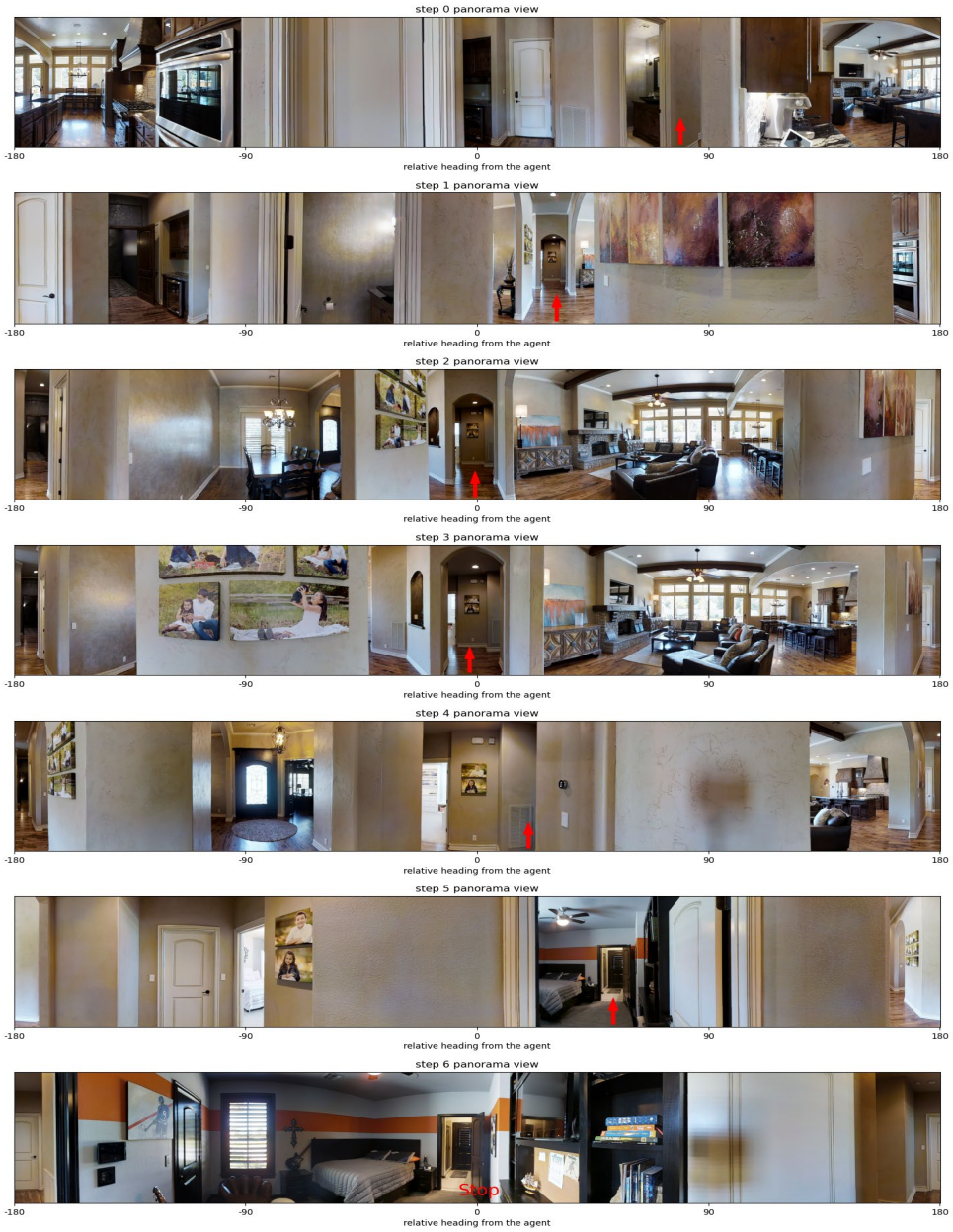
The **red** bounding box denotes the localized output object and the **green** bounding boxes are candidate objects.

Figure 3. The successful navigation and localization qualitative example result on Val Seen dataset.

Instruction

Leave the kitchen into the hall to your right and cross the hall to other end of house and to your right there is a bedroom with grey walls and a orange stripe going around the room and look at the black control on the wall to your left soon as you enter the bedroom

Navigation steps of the scene-intuitive agent. The **red** arrow shows the action chosen by the agent. The **Stop** sign means the agent stops at corresponding viewpoint.



The **red** bounding box denotes the localized output object and the **green** bounding boxes are candidate objects.

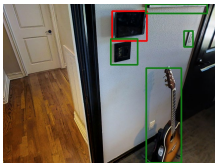


Figure 4. The successful navigation and localization qualitative example result on Val Seen dataset.

Instruction

Go to the mudroom and clean the counter

Navigation steps of the scene-intuitive agent. The **red** arrow shows the action chosen by the agent. The **Stop** sign means the agent stops at corresponding viewpoint.

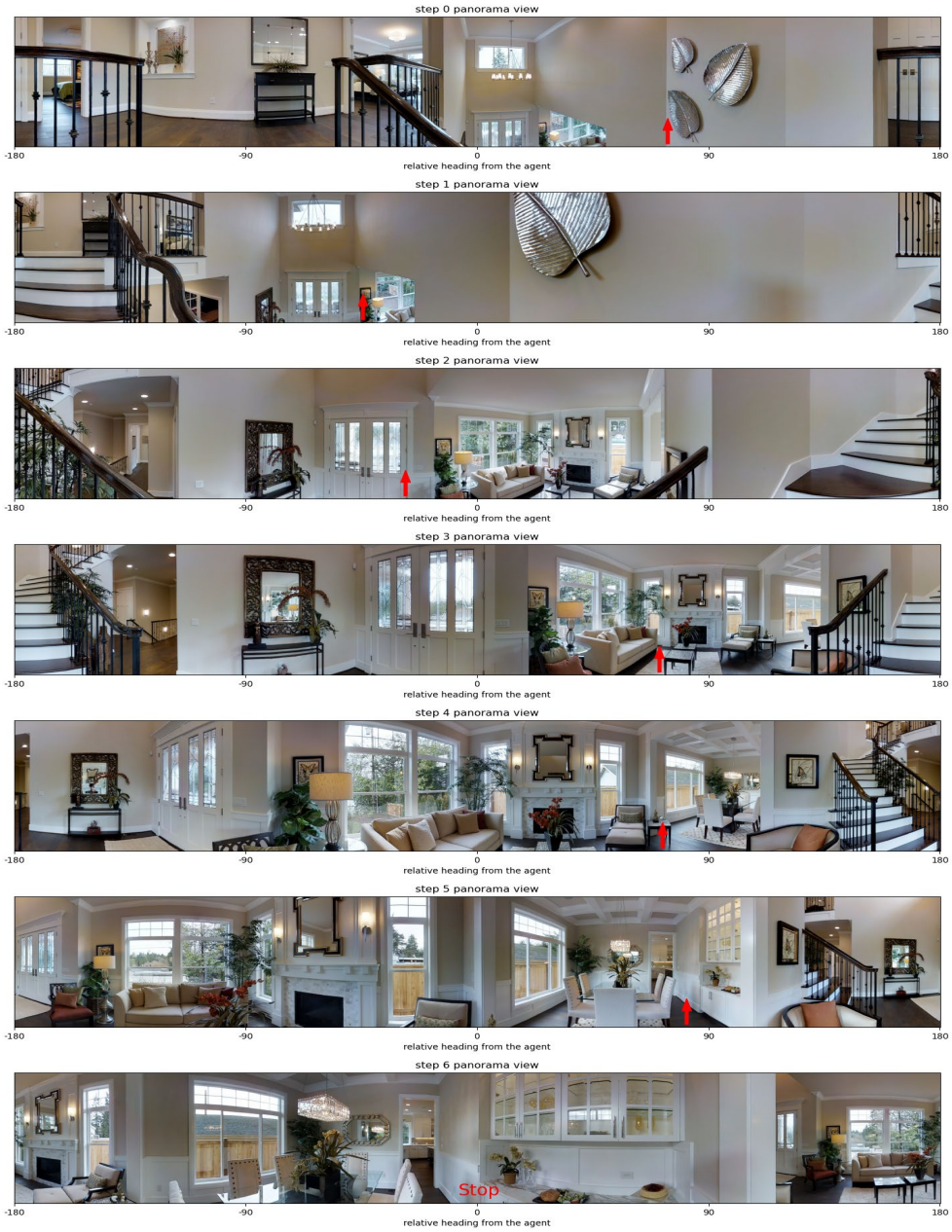
The **red** bounding box denotes the localized output object and the **green** bounding boxes are candidate objects.

Figure 5. The successful navigation and localization qualitative example result on Val Unseen dataset.

Instruction

Go to the dining room on level 2 and smash the table vase

Navigation steps of the scene-intuitive agent. The **red** arrow shows the action chosen by the agent. The **Stop** sign means the agent stops at corresponding viewpoint.



The **red** bounding box denotes the localized output object and the **green** bounding boxes are candidate objects.

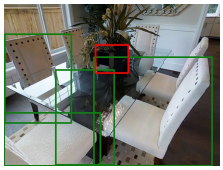
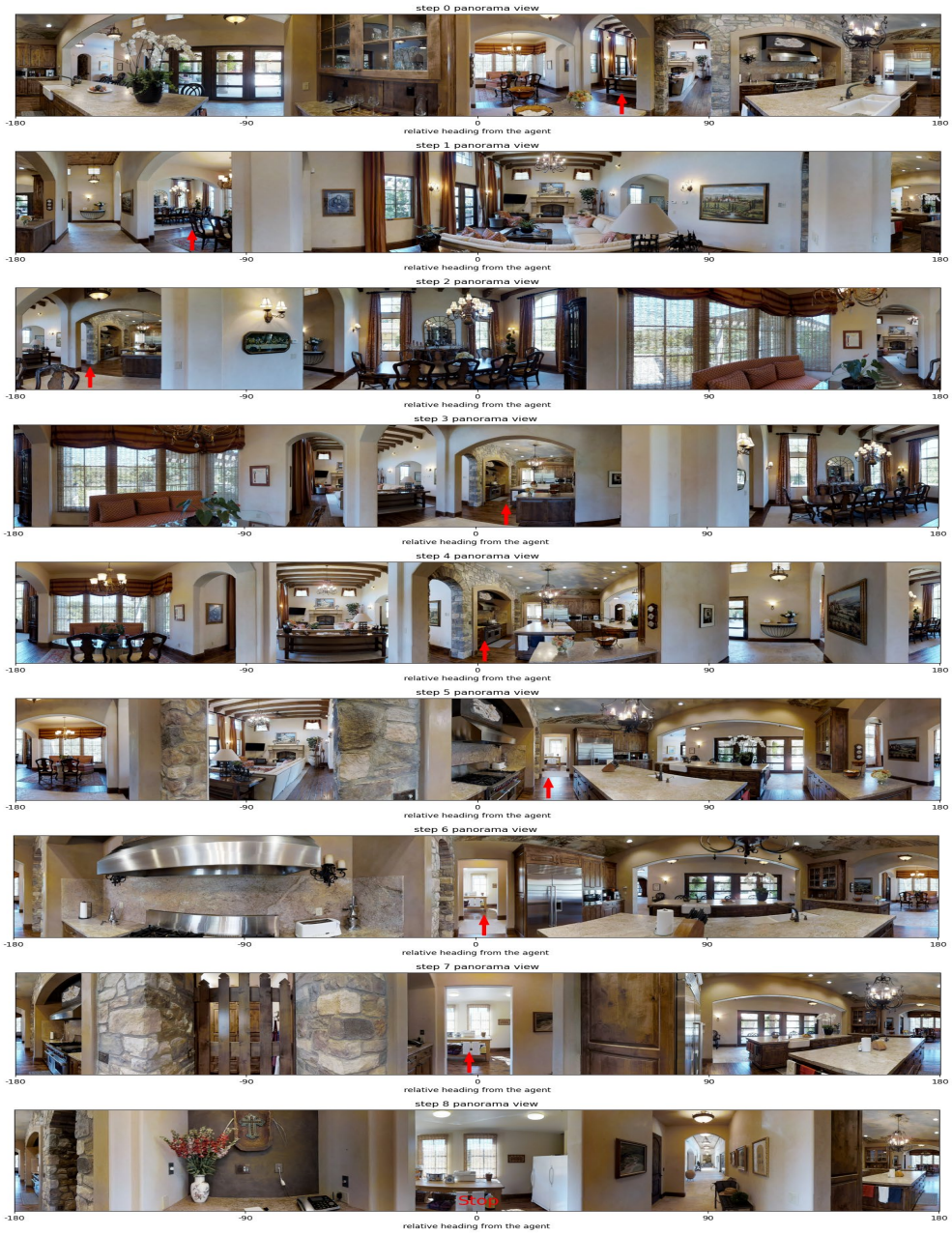


Figure 6. The successful navigation and localization qualitative example result on Val Unseen dataset.

Instruction

Go to the hallway on level 2 with the cross hanging on the wall and bring me the photo across from the light switch

Navigation steps of the scene-intuitive agent. The **red** arrow shows the action chosen by the agent. The **Stop** sign means the agent stops at corresponding viewpoint.



The **red** bounding box denotes the localized output object and the **green** bounding boxes are candidate objects.

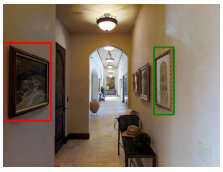
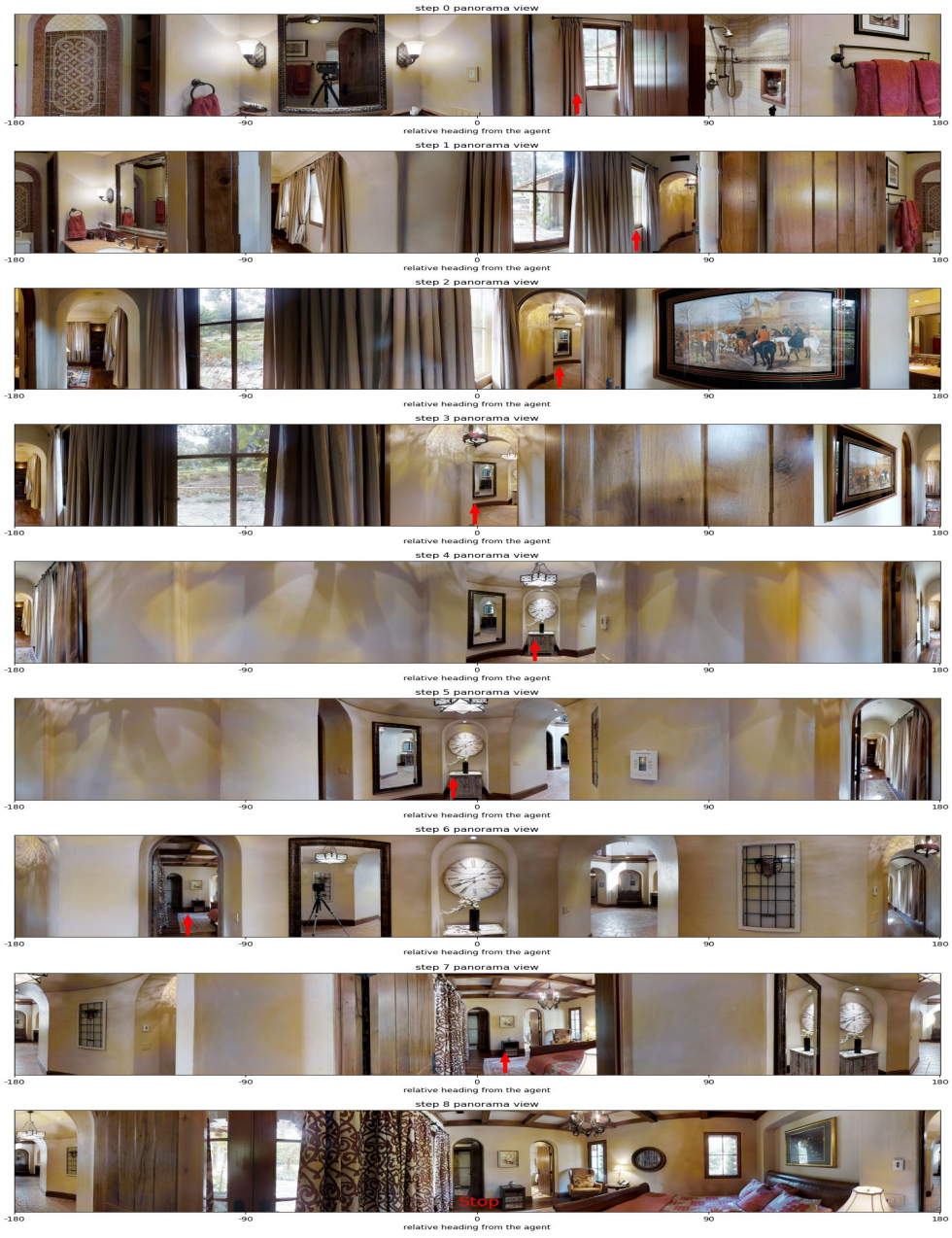


Figure 7. The successful navigation and localization qualitative example result on Val Unseen dataset.

Instruction

Go to the bedroom on level 1 that has a lot of red designs on top of the bed a painting depicting flowers above the bed and a ceiling-mounted chandelier in the center of the room and tell me if the chandelier is turned on

Navigation steps of the scene-intuitive agent. The **red** arrow shows the action chosen by the agent. The **Stop** sign means the agent stops at corresponding viewpoint.



The **red** bounding box denotes the localized output object and the **green** bounding boxes are candidate objects.

Figure 8. The successful navigation and localization qualitative example result on Val Unseen dataset.

Instruction

Go to second level balcony attached to the master bedroom and pick up the black and white stripped pillow closest to the tv

Navigation steps of the scene-intuitive agent. The **red** arrow shows the action chosen by the agent. The **Stop** sign means the agent stops at corresponding viewpoint.

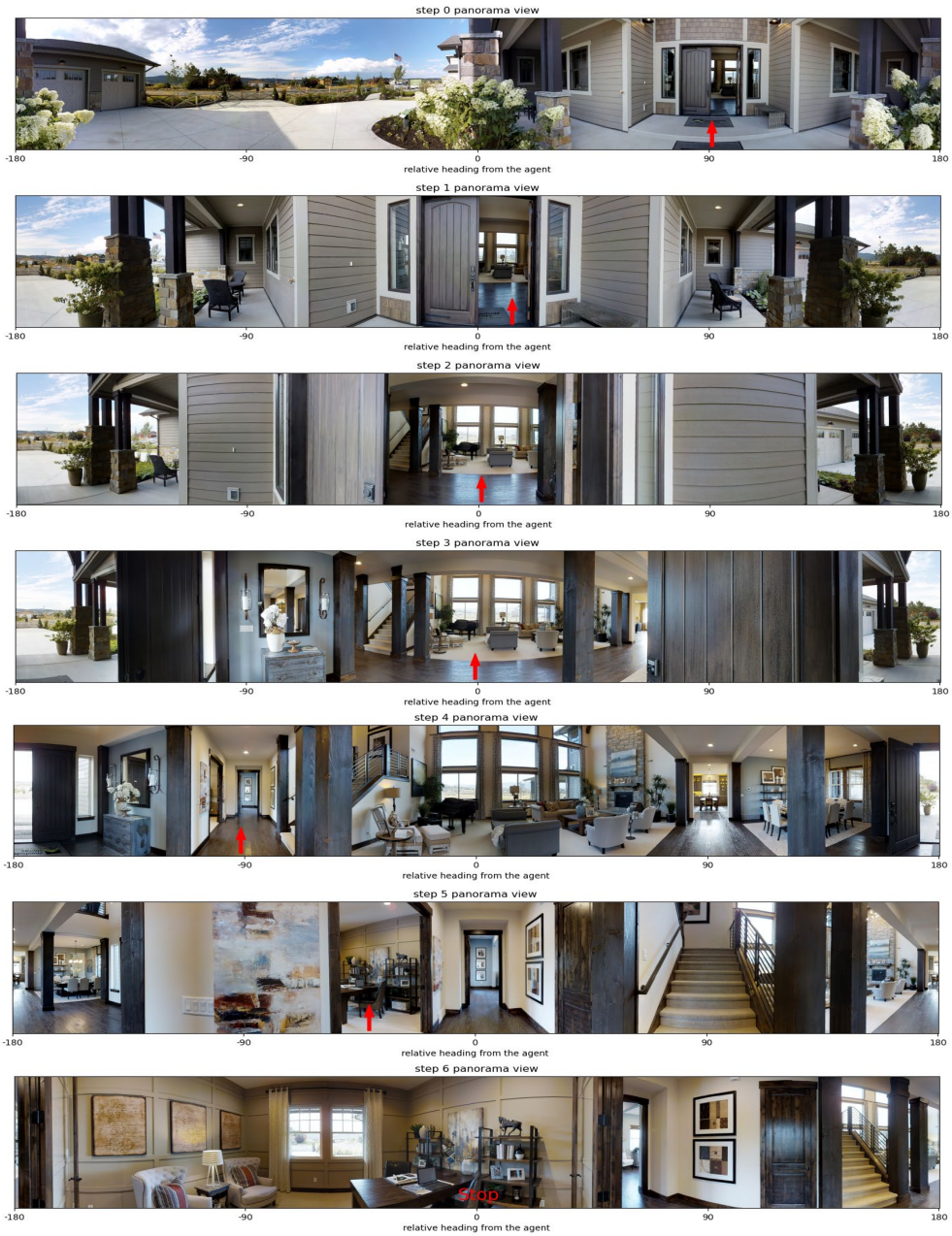
The left column denotes the agent's navigation steps and the right column shows the ground truth navigation steps.

Figure 9. The failed navigation qualitative example result. In this example, the agent first successfully navigates to second level but failed to enter the correct bedroom and stopped at a wrong viewpoint.

Instruction

Push in the chair nearest the door in the office

Navigation steps of the scene-intuitive agent. The **red** arrow shows the action chosen by the agent. The **Stop** sign means the agent stops at corresponding viewpoint.



The **red** bounding box denotes the localized output object, the **green** bounding boxes are candidate objects and the **blue** bounding box is the ground truth bounding box.

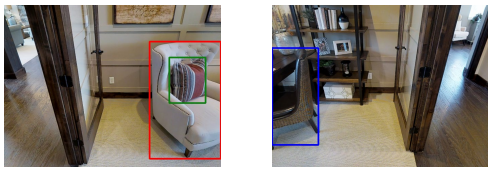


Figure 10. The failed localization qualitative example result. In this example, the agent first successfully navigates to the target viewpoint but failed to localize the target object because the $ViLPointer$ module thinks the white chair is closer to the office door than the black chair, which is reasonable as it is hard to decide which one is closer.

Instruction

Go to the dining room on level 2 and bring me the plate

Navigation steps of the scene-intuitive agent. The **red** arrow shows the action chosen by the agent. The **Stop** sign means the agent stops at corresponding viewpoint.

The **red** bounding box denotes the localized output object, the **green** bounding boxes are candidate objects and the **blue** bounding box is the ground truth bounding box.

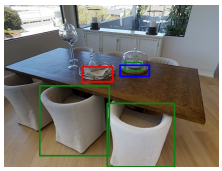
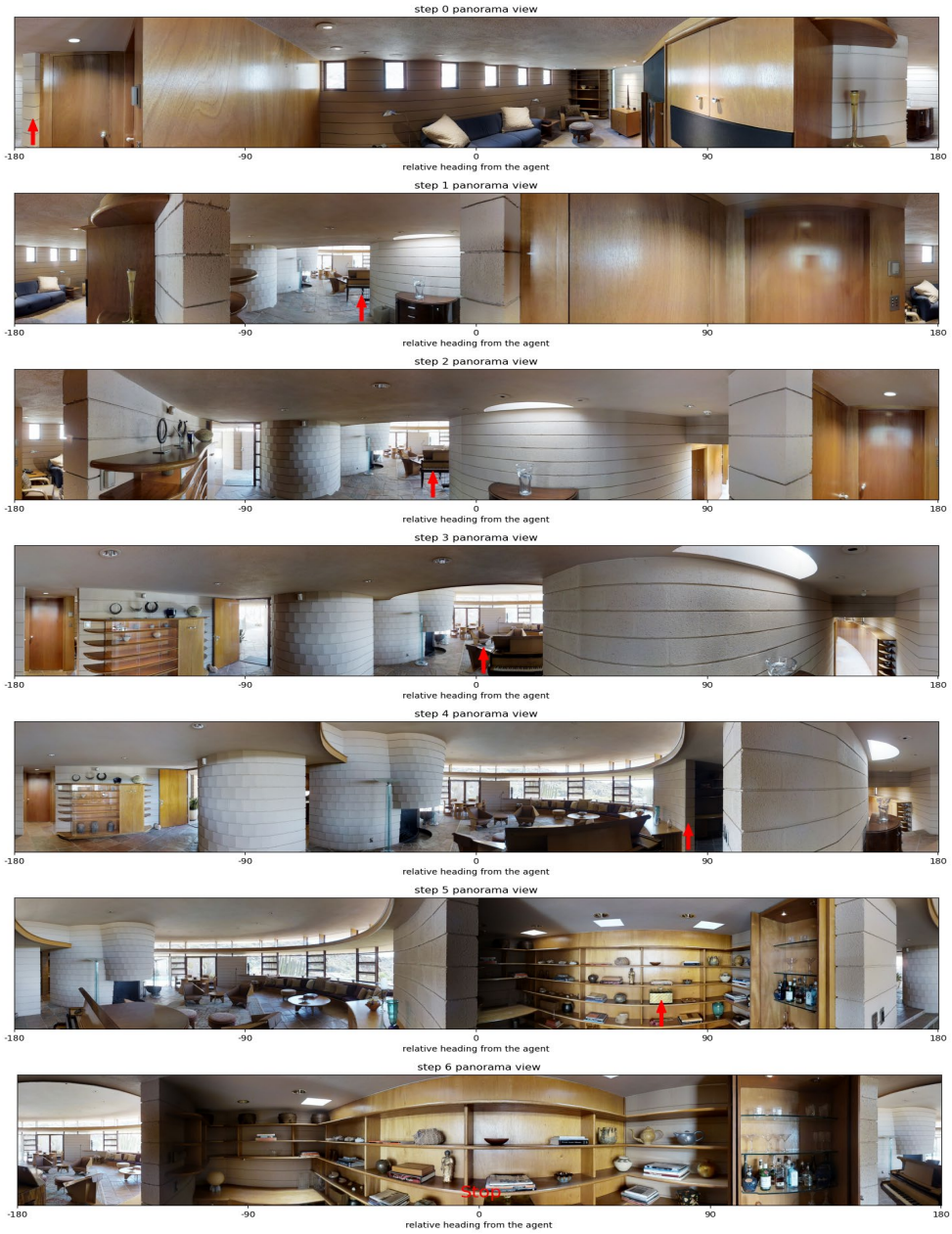


Figure 11. The failed localization qualitative example result. In this example, the agent first successfully navigates to the target viewpoint but failed to localize the target object because of the ambiguous meaning of "the plate" in the high-level instruction.

Instruction

Go to the office that is next to a piano on level 1 and bring the bottle from the shelf

Navigation steps of the scene-intuitive agent. The **red** arrow shows the action chosen by the agent. The **Stop** sign means the agent stops at corresponding viewpoint.



The **red** bounding box denotes the localized output object, the **green** bounding boxes are candidate objects and the **blue** bounding box is the ground truth bounding box.

Figure 12. The failed localization qualitative example result. In this example, the agent first successfully navigates to the target viewpoint but failed to localize the target object because of the ambiguous meaning of "the bottle on the shelf" in the high-level instruction.

Instruction

Go to second level balcony attached to the master bedroom and pick up the black and white stripped pillow closest to the tv

Navigation steps of the scene-intuitive agent. The **red** arrow shows the action chosen by the agent. The **Stop** sign means the agent stops at corresponding viewpoint.

The **red** bounding box denotes the localized output object, the **green** bounding boxes are candidate objects and the **blue** bounding box is the ground truth bounding box.

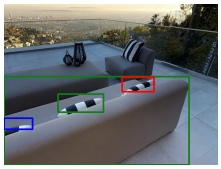


Figure 13. The failed localization qualitative example result. In this example, the agent first successfully navigates to the target viewpoint but failed to localize the target object because it failed to capture the relative position of similar objects ("the black and white stripped pillow") in the scene.