Supplementary Material
ADAPT: Vision-Language Navigation with Modality-Aligned Action Prompts

Bingqian Lin[1], Yi Zhu[2], Zicong Chen[1], Xiwen Liang[1], Jianzhuang Liu[2], Xiaodan Liang[1]
[1]Shenzhen Campus of Sun Yat-sen University [2]Huawei Noah's Ark Lab
{linbq6@mail2,chenzc7@mail2,liangxw29@mail2, liangxd9@mail}.sysu.edu.cn,
{zhuyi36, liu.jianzhuang}@huawei.com

A. Overview

In these supplementary materials, we first give more model details and implementation details in Sec. B and Sec. C, respectively. Then we present more quantitative results in Sec. D. More visualization results are given in Sec. E.

B. Model Details

The self-attention module $\mathrm{SelfAttn}(\cdot)$ and the cross-modal attention module $\mathrm{CrossAttn}(\cdot)$ are the conventional multi-head self-attention modules in the standard Transformer. Specifically, given the query Q, the key K, and the value V, the self-attention of the k-th attention head at the l-th layer is calculated as follows:

$$H_{l,k} = \mathrm{Softmax}(\frac{QK^{\mathrm{T}}}{\sqrt{d_h}})V, \tag{1}$$

where d_h is the hidden dimension of the network. The attention $H_{l,k}$ is used to update the query Q through the Feed-Forward Networks (FFN). In $\mathrm{SelfAttn}(\cdot)$, Q, K and V are obtained from the same single-modal feature. While in $\mathrm{CrossAttn}(\cdot)$, they are derived from the features of different modalities. The outputs of both modules are the self-attention values and the updated query features.

C. Implementation Details

Training. Following [2], we use a batch size of 16 during training for both R2R and RxR. We employ a two-stage training strategy for ADAPT, i.e., first train the baseline model [2] until the performance is converged in Val Unseen, and then pick the model with the highest SPL from the first stage and continue to train it with our ADAPT. The learning rates in the first and second stages are set to 1e-5 and 1e-6, respectively. The optimizer is AdamW [3]. The sequential consistency loss is used in both the imitation learning training and the reinforcement learning training, while the modality alignment loss is used only in the imitation learning training. During modality alignment learning, the non-paired sub-prompts for a specific sub-prompt are the sub-prompts in other samples in the same batch.

Action Prompts. Before feeding to the prompt encoder, the image and text sub-prompt features are extracted in advance through the visual and text encoders of CLIP [4], respectively. For accelerating the ADAPT training and inference, we perform the action prompt retrieval for each instruction in advance. The cosine similarity between two phrase features is calculated as the sentence similarity. The visual object/location vocabulary and the verb vocabulary are manually built by filtering the vocabulary of R2R given in [1]. To mitigate the multiple object noises existing in collecting the action prompts, we create a shared prompt base for R2R and RxR using the Fine-grained R2R[1], which contains paired sub-paths and sub-instructions always regarding single object/location. By selecting image sub-prompts from the sub-path rather than the whole path, our model can largely mitigate the noise.

D. More Quantitative Results

In this subsection, we present the quantitative comparison between some variants, i.e., "Object", "Extra", "Whole" and our ADAPT in Table 1. Specifically, "Object" means the paired text and image sub-prompts are replaced by object words and related object images, respectively. "Extra" means using action prompts as training samples directly. "Whole" means selecting image sub-prompts from the whole path rather than the sub-path in Fine-grained R2R. From the comparison results between "Object" and ADAPT we can observe that learning action-level alignment is more helpful for successful navigation than object-level alignment. The comparison results between "Extra" and ADAPT show the advantage of explicit action prompt learning over implicit training. The comparison results between "Whole" and ADAPT demonstrates that ADAPT effectively mitigates the multiple object noise during the ac-

[1]https://github.com/YicongHong/Fine-Grained-R2R

tion prompt collection.

E. More Visualization Results

Object-level Alignment vs. Action-level Alignment. Fig. 1 gives an action selection comparison between the models provided with object-level prompts and action-level prompts. From Fig. 1 we can observe that with action prompts indicating Enter the closet, ADAPT can successfully choose the action image which contains the closet door to enter it. The model provided with object prompts about the closet, however, selects the wrong action. These results show the necessity of action-level alignment.

Implicit Learning vs. Explicit Learning. We give a language attention comparison between the baseline agent [2] and ADAPT in Fig. 2, where we can find that both models attend to the right instruction part in steps 1-3. However, the baseline conducts wrong modality alignment and action. This shows that compared with attention-based implicit learning through simple navigation supervision, explicit action prompt learning contributes to learning better modality alignment for successful navigation.

Navigation Trajectories. Fig. 3 and Fig. 4 present examples of the panoramic views and action comparison between the baseline [2] and our ADAPT. From the visualization results we find that by introducing action prompts, ADAPT can make action decision accurately to accomplish successful navigation. For example, in Fig. 3, with the help of the action prompts related to "past the windows", ADAPT makes the correct action of "past the windows" in the first two navigation steps. The baseline agent, however, fails to conduct the action of "past the windows" during navigation and thus makes a wrong trajectory.

Failure Cases of Navigation Trajectories. Fig. 5 and Fig. 6 give failed trajectory examples of our ADAPT and the ground-truth. From the failure cases we can see that ADAPT may fail when the observations and instructions cause an ambiguity during navigation. From Fig. 5 we observe that although our ADAPT makes a wrong trajectory due to ambiguous observations of multiple "living area" and "wooden chair", it still conducts the correct actions of "go to the living area" and "wait at the wooden chair" in Step 2 and Step 4, respectively. From Fig. 6 we can find that due to the ambiguous instruction of "walk forward and turn left" without referring to concrete visual objects, our ADAPT makes the action of "turn left" before passing the stairs (the ground-truth one should be conducted after passing the stairs) and thus leads to a wrong trajectory. However, it still conducts the asked action of "entering the living room" at the end of the navigation.

Action Prompt Alignment. Fig. 7 presents additional results of action prompt alignment between the CLIP [4] features and the sub-prompt features of our ADAPT. For the action phrase feature, the top 5 similar image features are retrieved from the object/location-related image sub-prompt set. From Fig. 7 we can observe that compared with CLIP, our ADAPT can perform better action-level modality alignment. For example, in Fig. 7 (b) our ADAPT can effectively retrieve the closet images containing the appearances of the closet and its door through which the agent can make the action of "stop in front of".

Table 1. More quantitative comparison between some variants and our ADAPT.

Method	ResNet-152			CLIP		
	NE ↓	SR ↑	SPL	NE ↓	SR ↑	SPL ↑
Object	4.09	59.9	54.9	**4.01**	61.9	55.4
Extra	4.41	59.0	54.2	4.29	60.7	55.1
Whole	**4.05**	61.8	55.3	4.14	60.5	55.2
ADAPT	4.07	**62.5**	**56.1**	4.10	**63.1**	**57.2**

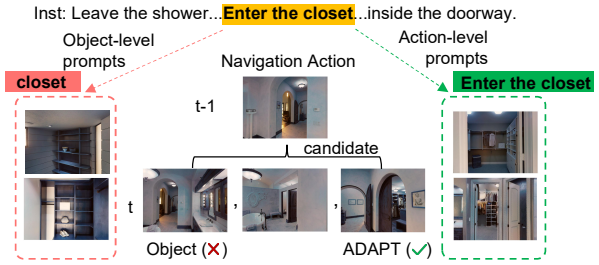

Figure 1. Action selection comparison between the models provided with object-level prompts and action-level prompts.

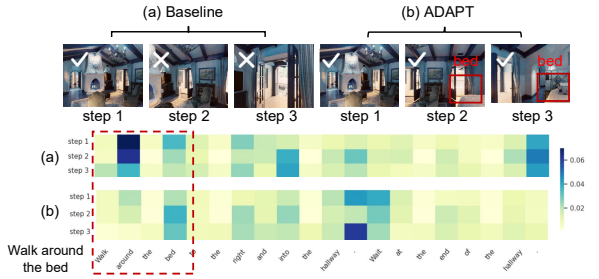

Figure 2. Language attention comparison between the baseline [2] and ADAPT.

References

[1] Peter Anderson, Qi Wu, Damien Teney, Jake Bruce, Mark Johnson, Niko Sunderhauf, Ian Reid, Stephen Gould, and Anton 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 (CVPR)*, pages 3674–3683, 2018. 1

[2] Yicong Hong, Qi Wu, Yuankai Qi, Cristian Rodriguez-Opazo, and Stephen Gould. Vln bert: A recurrent vision-and-language bert for navigation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1643–1653, 2021. 1, 2, 4

[3] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2018. 1

[4] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the International Conference on Machine Learning (ICML)*, pages 8748–8763, 2021. 1, 2

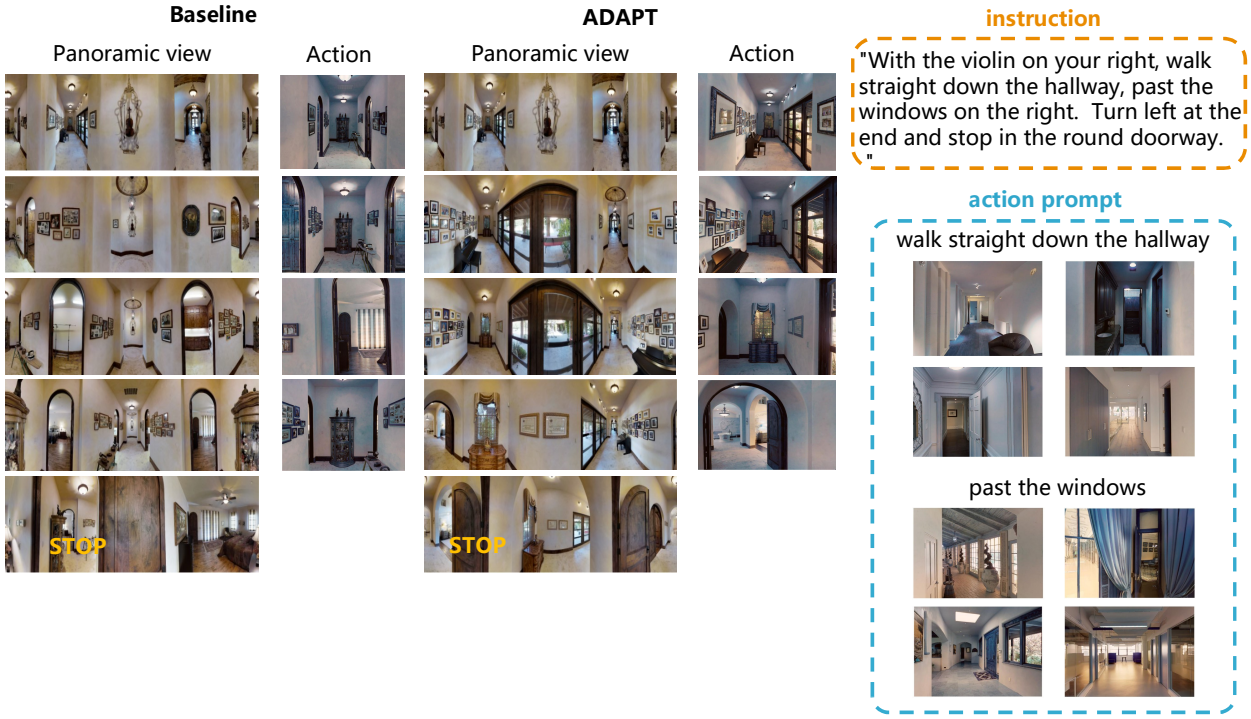

Figure 3. Visualization of panoramic views and action comparison in a trajectory example between the baseline [2] and our ADAPT.

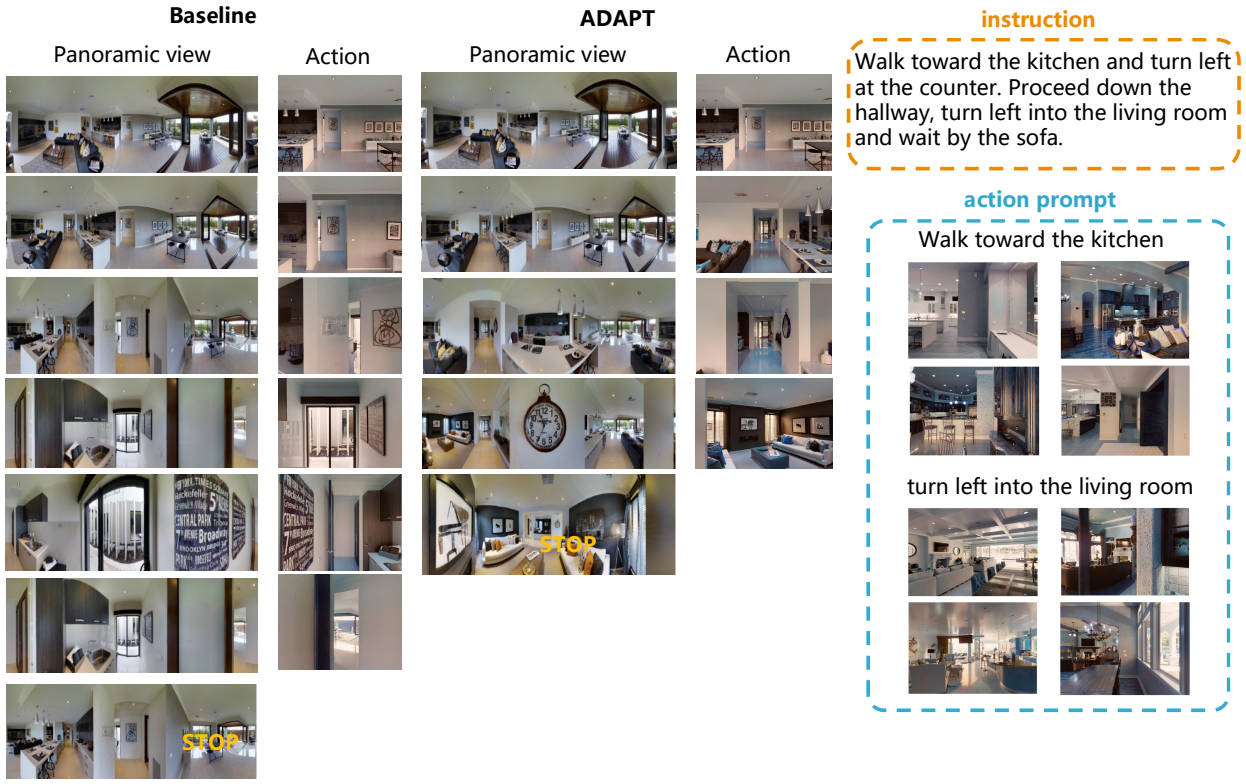

Figure 4. Visualization of panoramic views and action comparison in a trajectory example between the baseline [2] and our ADAPT.

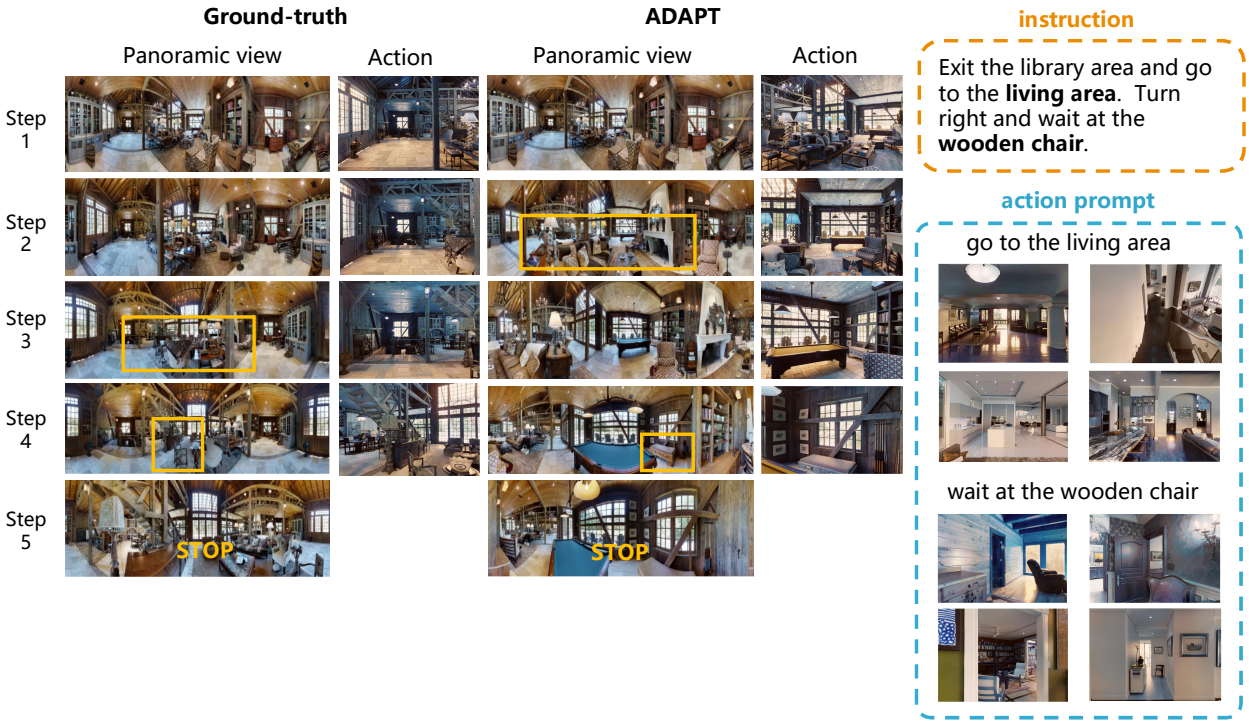

Figure 5. A failed trajectory example of our ADAPT and the related ground-truth. The yellow boxes indicate the instruction-related visual objects/locations appearing during the navigation trajectory.

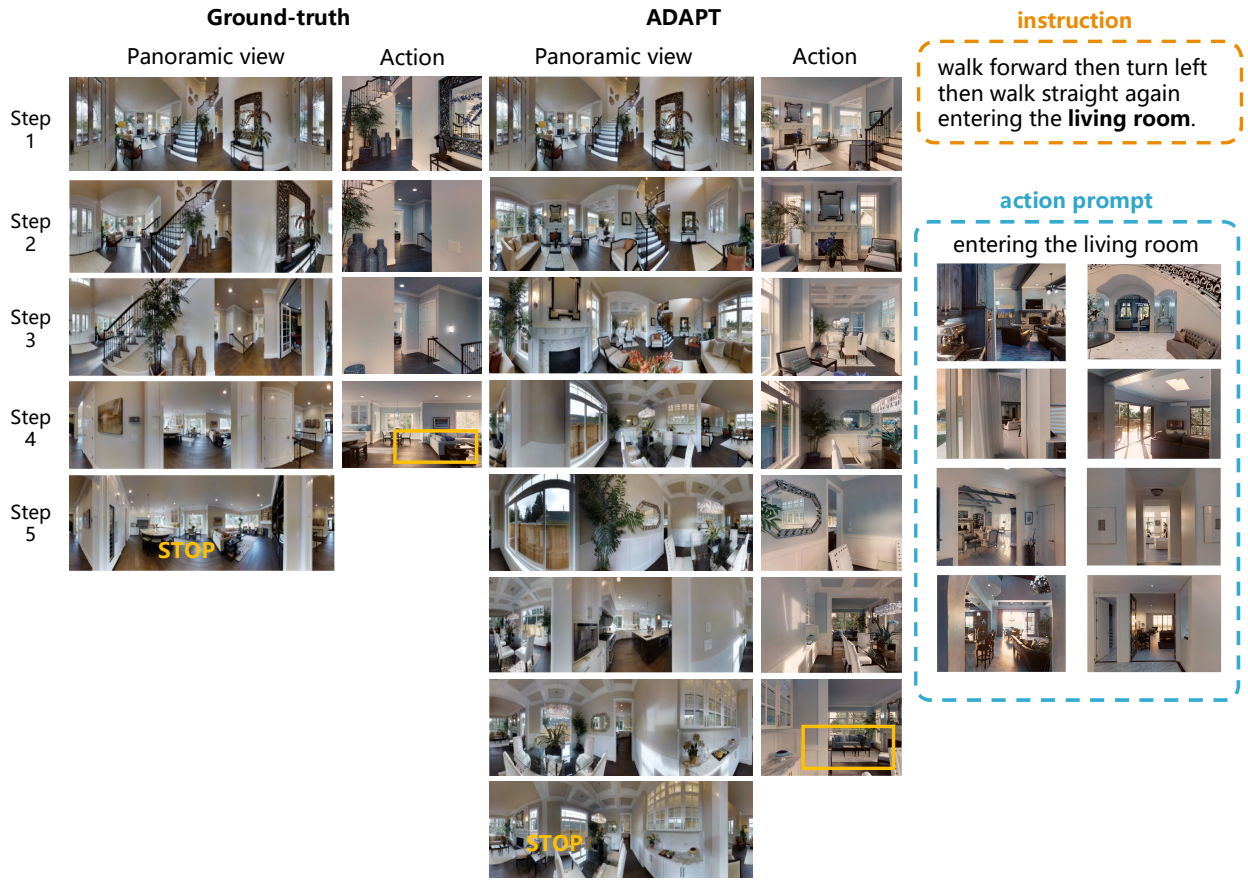

Figure 6. A failed trajectory example of our ADAPT and the related ground-truth. The yellow boxes indicate the instruction-related visual objects/locations appearing during the navigation trajectory.

Figure 7. Action prompt alignment comparison between the CLIP features and the sub-prompt features of our ADAPT. "✓": correct; "✕": incorrect; "**?**": ambiguous.