Supplementary Material for "Learning from Unlabeled 3D Environments for Vision-and-Language Navigation"

Shizhe Chen[1], Pierre-Louis Guhur[1], Makarand Tapaswi[2], Cordelia Schmid[1], and Ivan Laptev[1]

[1] Inria, École normale supérieure, CNRS, PSL Research University
[2] IIIT Hyderabad

We first present additional automatically generated examples from our HM3D-AutoVLN dataset in Sec. A. Then, we present further ablation analysis of our methods and also discuss results on the SOON dataset (equivalent analysis for the REVERIE dataset was in the main paper) in Sec. B. Finally, some qualitative visualizations of predicted navigation results are presented in Sec. C.

A HM3D-AutoVLN Dataset

In Fig. 1, we provide more examples of generated navigation graphs and object-trajectory-instruction triplets in the HM3D-AutoVLN dataset. Each trajectory consists of a sequence of panoramas. We only visualize the oriented view image at each step in Fig. 1 for simplicity.

B Additional Ablations

B.1 Comparison of Different Generated Instructions

We compare our generated instructions from the proposed speaker model (Sec. 3.2 of the main paper) with two types of template-based instructions. The first type of template instructions (Template.Obj) only describes the target object name, *e.g.*, "open the window", "clean the mirror". The second type (Template.Sent) uses sentence structures in REVERIE training split as templates [1] and replaces the noun phrases in the template with detected objects in the scene, *e.g.*, "go to bedroom and clean the chandelier". As these template instructions are less accurate and contain less details, we generate three instructions per target object to alleviate the impact of noisy instructions. As shown in Table 1, even with fewer number of instructions per object, the generated instructions from our speaker model outperform the two template-based instructions.

B.2 Influence of the Number of Training Environments

Table 2 and Table 3 present the results for all evaluation metrics corresponding to the SR/SPL results in Fig. 4 of the main paper. We can see that more training environments are beneficial for the performance on unseen environments in

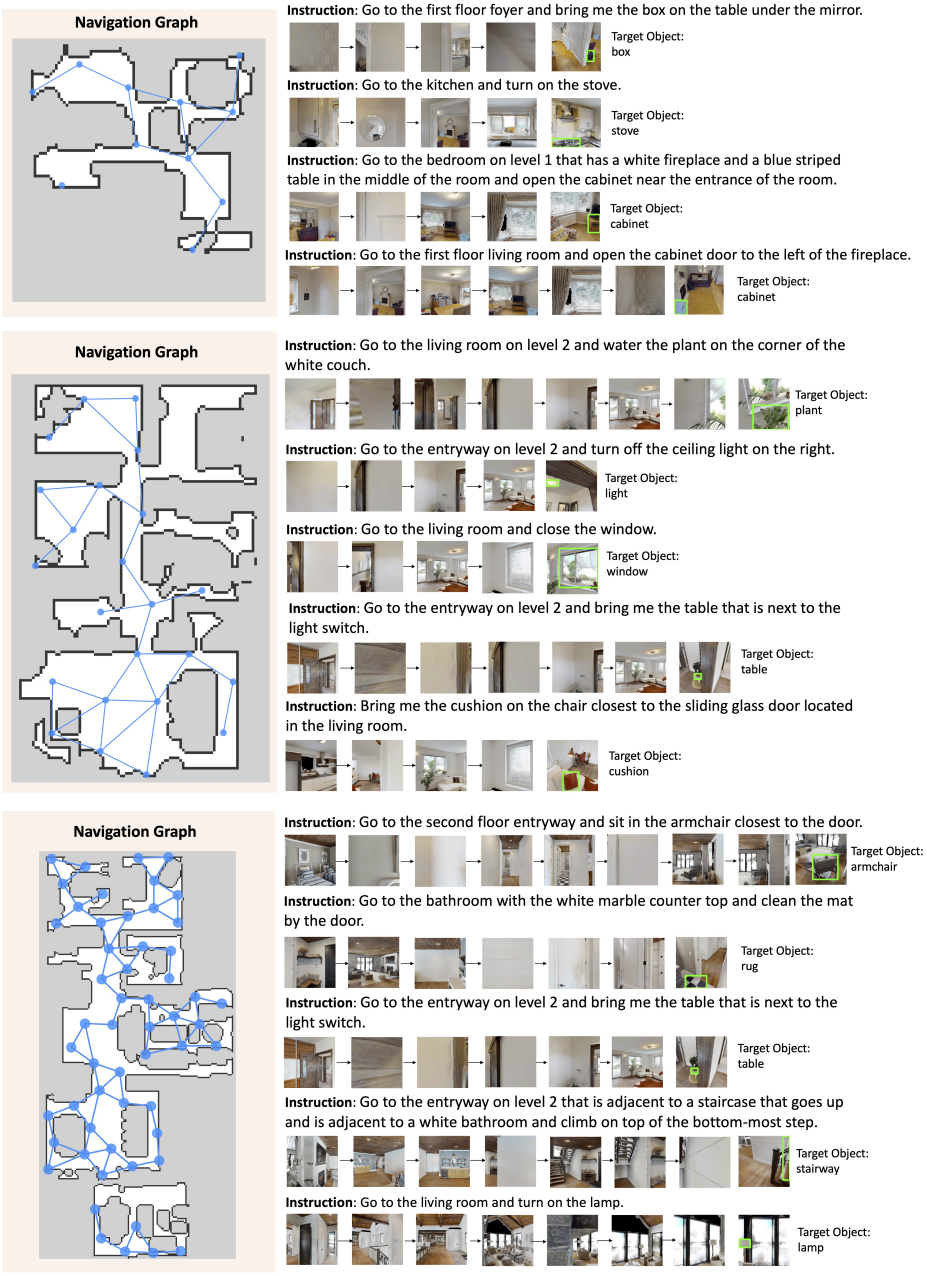

Fig. 1: Examples of generated navigation graphs and object-trajectory-instruction triplets in HM3D-AutoVLN dataset.

Table 1: DUET performance on the REVERIE val unseen split using different methods for generating instructions.

Method	#Instrs	inst.len	Navigation			Grounding	
			OSR	SR	SPL	RGS	RGSPL
Template.Obj	653,109	3.88	54.33	49.28	37.66	29.85	23.07
Template.Sent	653,109	11.50	56.16	52.20	39.90	32.75	24.86
GPT2 (Ours)	217,703	20.52	**62.14**	**55.89**	**40.85**	**36.58**	**26.76**

Table 2: DUET performance on the REVERIE val unseen split using different number of training environments from the HM3D-AutoVLN dataset.

#Envs	Navigation			Grounding	
	OSR	SR	SPL	RGS	RGSPL
200	55.61	50.13	37.65	32.92	24.68
400	55.04	51.09	36.79	32.43	23.16
600	60.61	53.99	40.40	33.85	25.52
900	**62.14**	**55.89**	**40.85**	**36.58**	**26.76**

Table 2 and that this result is reflected across all metrics. When using a fixed number of instructions, it is better to collect training examples from more environments as shown in Table 3.

B.3 Few-shot Results

In Table 4, we show the performance of using different amounts of training data from the SOON dataset to fine-tune the DUET model pretrained on our HM3D-AutoVLN dataset. This is comparable to Table 3 of the main paper that presented results on the REVERIE dataset. To be noted, instructions in SOON dataset are not used to train our speaker model, and the style of instructions in HM3D-AutoVLN is quite different from the style in the SOON dataset. Despite the style differences, the pretraining allows our model to achieve better performance with a small amount of supervised data.

C Qualitative Results

Fig. 2 compares some predicted trajectories from DUET models with and without pretraining on our HM3D-AutoVLN dataset. We can see that the pretraining allows the navigation model to explore the unseen environment more efficiently. Fig. 3 further presents some failure cases. The main problem is to correctly decide the stop location aligned to the fine-grained instruction. For example, in the last two examples in Fig. 3, the agent has navigated close to the target location but failed to stop and continued the exploration.

Table 3: DUET performance on the REVERIE val unseen split using the same number of instructions but different number of training environments in HM3D-AutoVLN dataset (51,018 is the total number of instructions contained in the 200 training environments, and so is 99,224 in 400 training environments)

#Instrs	#Envs	Navigation			Grounding	
		OSR	SR	SPL	RGS	RGSPL
51,018	200	55.61	50.13	37.65	32.92	24.68
	900	54.05	50.75	39.52	32.35	25.22
99,224	400	55.04	51.09	36.79	32.43	23.16
	900	60.44	53.96	38.35	36.01	25.56

Table 4: DUET performance on SOON val unseen split using different number of SOON training environments in fine-tuning

Pretrain	#Envs	#Instrs	Navigation			Grounding
			OSR	SR	SPL	RGSPL
×	34	2,780	50.91	36.28	22.58	3.75
✓	0	0	14.75	11.21	7.92	0.38
✓	1	67	37.32	25.22	17.97	2.17
✓	10	848	45.58	34.51	24.38	2.75
✓	34	2,780	53.19	41.00	30.69	4.06

References

1. Guhur, P.L., Tapaswi, M., Chen, S., Laptev, I., Schmid, C.: Airbert: In-domain pretraining for vision-and-language navigation. In: ICCV. pp. 1634–1643 (2021) 1

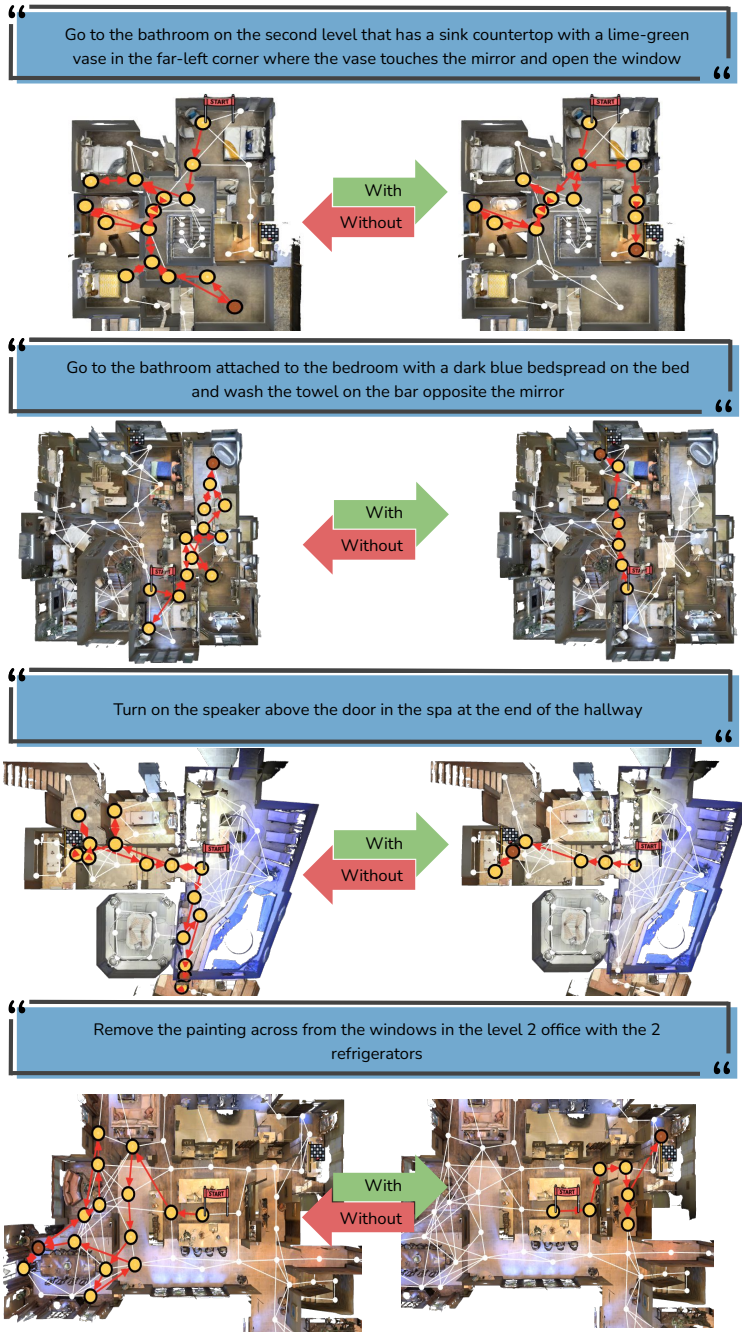

Fig. 2: Examples where pretraining with HM3D-AutoVLN improved the performance of DUET on REVERIE val unseen split. The checkered flag and maroon node denote the target and predicted destinations respectively.

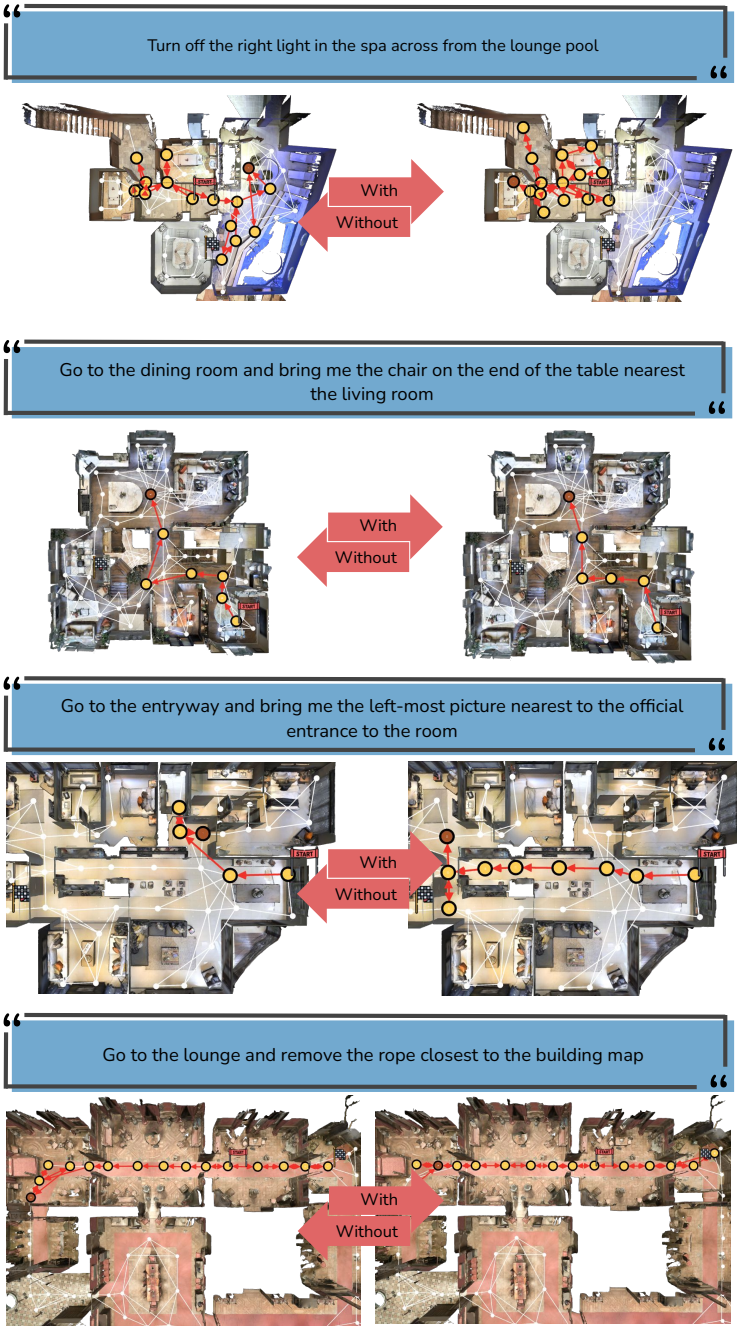

Fig. 3: Examples where DUET models with and without pretraining on HM3D-AutoVLN both fail on REVERIE val unseen split. The checkered flag and maroon node denote the target and predicted destinations respectively.