Controllable Navigation Instruction Generation with Chain of Thought Prompting

Supplementary Material

This document provides more details, extra experimental results, and further discussion of C-Instructor. The document is organized as follows:

- §A provides detailed prompts for several datasets.
- §B presents extra ablation results on different selection strategies and values of β in landmark selection.
- §C further analyzes the effect of STMT through the training process.
- §D shows more qualitative results of instruction generation and analyzes some failure cases.
- §E discusses the social impact and limitations of our work, and suggests potential future work.

A Detailed Prompts

In this section, we provide detailed prompts for different navigation datasets. Note that all the given prompts are then formatted by prompt templates in [2].

- $\texttt{prompt}_{\lambda}$ for R2R [1]: You are given a sequence of views of a path. Please extract critical landmarks in the path.
- $\texttt{prompt}_{w}$ for R2R [1]: You are given a sequence of views of a path in an indoor environment. Please describe the path according to the given landmarks in detail for an intelligent agent to follow. Landmarks: <*landmarks*>.
- $\texttt{prompt}_{\lambda}$ for REVERIE [4]: You are given a sequence of views of a path in an indoor environment. Please extract several critical landmarks in the path for generating a brief high-level target-oriented instruction.
- $\texttt{prompt}_{w}$ for REVERIE [4]: You are given a sequence of views of a path in an indoor environment and critical landmarks for a brief high-level target-oriented instruction. Please generate the indicated high-level target-oriented instruction briefly for an intelligent agent to follow. Landmarks: <*landmarks*>.
- $\texttt{prompt}_{\lambda}$ for RxR [3]: You are given a sequence of views of a path in an indoor environment. Please extract critical landmarks describing the starting position and the path.
- $\texttt{prompt}_{w}$ for RxR [3]: You are given a sequence of views of a path in an indoor environment. Please describe the starting position and the path according to the given landmarks in detail for an intelligent agent to follow. Landmarks: <*landmarks*>.
- $\texttt{prompt}_{a}$: You are an intelligent embodied agent that navigates in an indoor environment. Your task is to move among the static viewpoints (positions) of a pre-defined graph of the environment. You are given several candidate views. You are also given a sequence of panoramic views showing previous

Table 1: Ablations on landmark selection strategies (§B.1) on REVERIE [4] `val unseen` and R2R [1] `val unseen`.

Methods	REVERIE `val unseen`						R2R `val unseen`					
	SPICE↑	BLEU-1↑	BLEU-4↑	CIDEr↑	Meteor↑	Rouge↑	SPICE↑	BLEU-1↑	BLEU-4↑	CIDEr↑	Meteor↑	Rouge↑
Baseline	0.129	0.737	0.402	0.490	0.258	0.590	0.194	0.689	0.262	0.445	0.228	**0.479**
Baseline + Λ_x	0.143	0.732	0.380	0.482	0.263	0.580	0.199	0.687	0.252	0.416	0.230	0.466
Baseline + $\Lambda_x \cup \Lambda_a$	**0.150**	0.748	0.401	0.531	0.263	0.583	0.207	0.707	0.250	0.424	0.232	0.466
Baseline + $\Lambda_x \cup \Lambda_v$	0.141	**0.754**	**0.419**	**0.545**	**0.267**	**0.591**	**0.212**	**0.713**	**0.266**	**0.447**	**0.239**	0.473

Table 2: Ablations on the value of β in landmark selection (§B.2) on REVERIE [4] `val unseen` and R2R [1] `val unseen`.

β	REVERIE `val unseen`						R2R `val unseen`					
	SPICE ↑	BLEU-1 ↑	BLEU-4 ↑	CIDEr ↑	Meteor ↑	Rouge ↑	SPICE ↑	BLEU-1 ↑	BLEU-4 ↑	CIDEr ↑	Meteor ↑	Rouge ↑
0	**0.150**	0.749	0.409	0.538	0.267	0.587	0.208	**0.719**	0.266	0.413	0.236	0.469
0.25	0.141	**0.754**	**0.419**	**0.545**	**0.267**	**0.591**	**0.212**	0.713	**0.266**	**0.447**	**0.239**	**0.473**
0.5	0.137	0.717	0.376	0.488	0.262	0.576	0.206	0.692	0.247	0.410	0.231	0.461

steps you have taken and the previous viewpoint you should return to. Now you should make an action by selecting a candidate view to return to the previous viewpoint. Candidate Views: <*viewpoints*>

B Extra Ablations on Landmark Selection

B.1 Selection Strategies

To validate the effectiveness of our landmark selection strategy, we conducted several experiments with several ablative strategies on REVERIE [4] and R2R [1] `val unseen` splits. The results are shown in Tab. 1.

#1 is the baseline result without landmarks and the CoT process. The model in #2 uses only landmarks from instructions Λ_x in CoTL. Compared to #1, the `SPICE` metric remarkably increases, which indicates a more accurate description of object relations in the instructions. Other metrics fluctuate. Based on #2, the model in #3 adds visual landmarks via spatial selection, which are denoted as Λ_a. Compared to #2, almost all metrics rise, which demonstrates the value of visual landmarks. The model in #4 adds visual landmarks via spatial and temporal selection Λ_v in addition to landmarks from instructions, resulting in an increase in almost all scores compared to #3. The results above further confirm the effectiveness of the proposed landmark selection mechanism.

B.2 Values of β

We conduct ablations on the numerical values of β in landmark selection on REVERIE [4] and R2R [1] `val unseen` splits. The results are presented in Tab. 2. It can be observed that assigning β to 0.25 achieves the best performance.

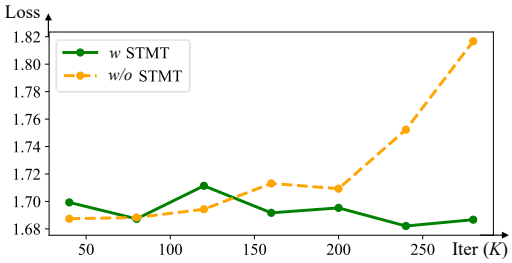


Fig. 1: Validation loss on R2R [1] `val unseen` (§C).

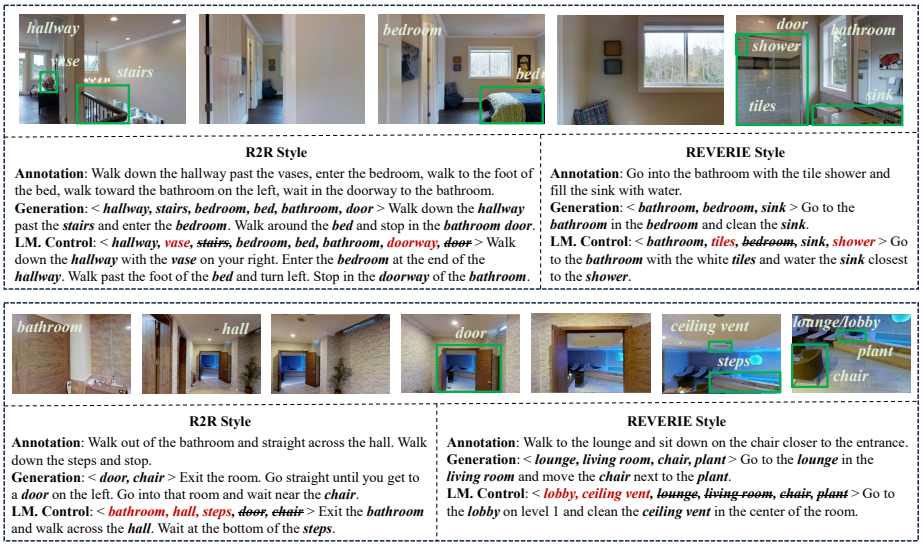


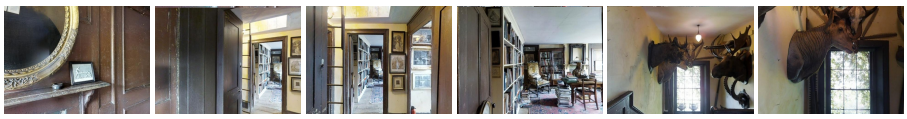
Fig. 2: Additional visualizations of navigation trajectories and instruction generation results on R2R [1] and REVERIE [4] (§D).

C Further Analysis on STMT

In Fig. 1, we plot the curve illustrating the model's validation loss on R2R [1] `val unseen` during the training process. Compared to the baseline without STMT, We can observe that STMT effectively prevents overfitting as evidenced by the fact that its validation loss does not exhibit a gradual increase compared to the baseline. STMT effectively ensures the training stability of C-Instructor and also enhances the instruction quality.

D Additional Qualitative Results

In Fig. 2, we provide more visualizations of navigation trajectories and corresponding instruction generation results. As observed, C-Instructor effectively



R2R Style	REVERIE Style
Annotation: Turn right and exit the bedroom. Turn left and walk through the doorway. Wait at the top of the stairs. **Generation**: Exit the bedroom and walk through the hallway to the stairs. Wait at the top of the stairs.	**Annotation**: Go to the stairs of ***level 3*** and wipe down the ***hunting trophy***. **Generation**: Go to the hallway on ***level 2*** and take down the ***picture*** closest to the stairs.

Fig. 3: Failure case of C-Instructor (§D).

identifies essential landmarks in the trajectory and generates high-quality instructions accordingly in specified linguistic styles. Control over the focus of C-Instructor can be achieved by manipulating landmarks. Modifying either part of landmarks (Fig. 2 upper) or all the landmarks (Fig. 2 lower) leads to reasonable instruction generation results.

Failure Case. We present a failure case of C-Instructor in Fig. 3. In this case, C-Instructor mistakes *level 3* as *level 2* for lack of knowledge of the global structure of the house. Furthermore, it misidentifies the rarely-seen object *hunting trophy* as a *picture*. This case suggests future efforts on global environmental structure encoding and more accurate object identification.

E More Discussion

Social Impact. C-Instructor can be used to provide feedback from intelligent embodied agents to humans as well as to guide humans who are unfamiliar with the environment. It can also serve as accessibility facilities for the visually impaired to find their way.

Limitations. Due to data availability, C-Instructor is trained on simulated data with discrete viewpoints, which limits its performance in real-world continuous environments. Moreover, as discussed in §D, C-Instructor possesses limited ability in modeling the global structure of the environment, resulting in inaccurate instructions when referring to the global location of a specific object or room in the environment.

Future Work. We plan to devise a mechanism that encodes the global structure of the environment into the instruction generator. With knowledge of the environment, the instruction generator can locate the user according to free-form natural language descriptions and provide path guidance according to the destination designated by the user.

References

1. Anderson, P., Wu, Q., Teney, D., Bruce, J., Johnson, M., Sünderhauf, N., Reid, I., Gould, S., Van Den Hengel, A.: Vision-and-language navigation: Interpreting visually-grounded navigation instructions in real environments. In: CVPR (2018)

2. Gao, P., Han, J., Zhang, R., Lin, Z., Geng, S., Zhou, A., Zhang, W., Lu, P., He, C., Yue, X., et al.: Llama-adapter v2: Parameter-efficient visual instruction model. arXiv preprint arXiv:2304.15010 (2023)
3. Ku, A., Anderson, P., Patel, R., Ie, E., Baldridge, J.: Room-across-room: Multilingual vision-and-language navigation with dense spatiotemporal grounding. In: EMNLP (2020)
4. Qi, Y., Wu, Q., Anderson, P., Wang, X., Wang, W.Y., Shen, C., Hengel, A.v.d.: Reverie: Remote embodied visual referring expression in real indoor environments. In: CVPR (2020)