Supplementary for NavGPT-2: Unleashing Navigational Reasoning Capability for Large Vision-Language Models

Gengze Zhou[1], Yicong Hong[2], Zun Wang[3], Xin Eric Wang[4], and Qi Wu[1(✉)]

[1] AIML, University of Adelaide, Adelaide, Australia
{gengze.zhou, qi.wu01}@adelaide.edu.au
[2] Adobe Research, San Jose, USA
[3] Shanghai AI Laboratory, Shanghai, China
[4] University of California, Santa Cruz, USA

Section 1 provides additional details for DUET as it is the main comparison with NavGPT-2. The prompt for GPT-4V used in the data generation pipeline and additional experiment results are described in Section 2 and Section 3. Section 4 illustrates the limitation of NavGPT-2 with the discussion of future directions. Finally, Section 5 discusses the broader impacts of our work.

1 DUET Revisit

NavGPT-2 exploit the similar design adapted from Dual-scale Graph Transformer (DUET) [1] as the downstream navigation policy. It includes a text encoder to encode instructions, a global and a local branch to enable coarse-scale and fine-scale cross-modal reasoning.

1.1 Text Embedding and Visual Embedding

For the text encoder, DUET utilizes a 12-layer transformer initialized from LXMERT [3]. For visual embedding, the visual observation at each node is 36 view images from 12 horizontal directions times 3 vertical directions. To distinguish these nodes, a directional embedding E^{ang} of the absolute angle for each view is added to the visual feature $\mathcal{Z}^v$ extracted by the vision encoder. Moreover, since DUET inputs all 36 view images to construct the spatial observation for the model, the navigable adjacent nodes are only observed at a few view images, denoted as navigable views. A navigable embedding E^{nav} is added to the visual features. The final visual embedding is sent to a 2 layers transformer to encode the spatial relations between views and obtain the panoramic view embeddings:

$$\mathcal{H}^{pano} = \text{SelfAttn}\left(\mathcal{Z}^v + E^{ang} + E^{nav}\right). \tag{1}$$

On the contrary, NavGPT-2 only inputs the navigable views, thus the directional embedding E^{ang} and the navigable embedding E^{nav} are removed in the downstream policy, instead we directly add the step embedding and location embedding before sending to the 2 layers transformer.

1.2 DUET Local Branch

NavGPT-2 adopt the same navigation policy network architecture as the DUET global branch, discussed in §3.2[5], so we omit the explanation of the global branch in DUET. In this section, we introduce the local branch of DUET. This branch performs action prediction based on the current node's instruction and egocentric observation. No graph information is provided besides the local observation.

Local Visual Embedding Two types of location embedding are added to the panoramic view embedding $\mathcal{H}^{pano}$. The first type is the relative location of the current node to the starting node, to encode the long distance direction between nodes. The second type is each adjacent view to the current node, to encode egocentric directions such as "turn right".

Local Cross-model Encoding The local branch utilizes a standard cross-modal transformer of 4 layers to model vision and language relations. During action prediction, a mask is set to the unnavigable views, and the action logits are only calculated for the navigable views at the current node.

1.3 Dynamic Fusion

The final action prediction of DUET is performed by dynamically fusing the action predicted by local and global branches. The local branch predicts actions within the adjacent nodes $\mathcal{V}_t^a$. It is incongruent with the action space used by the global branch, which chooses the next action from all nodes $\mathcal{V}_t$ in the constructed graph at step t. To reconcile this discrepancy, the local action scores s_i^l are transformed encompassing options such as "stop" and $\mathcal{V}_t$, into a representation suitable for the global action space by summing up scores of visited nodes in $\mathcal{V}_t^a$ as a backtrack score s_b:

$$s_i^{l'} = \begin{cases} s_{\mathrm{b}}, & \text{if } \mathcal{V}_i \in \mathcal{V}_t - \mathcal{V}_t^a, \\ s_i^{l'}, & \text{otherwise.} \end{cases} \tag{2}$$

This adjustment facilitates navigation toward other unexplored nodes not directly linked to the current node, necessitating the agent to retrace its steps through neighboring nodes that have previously been visited. The final navigation score is given by:

$$s_i = \sigma_t s_i^g + (1 - \sigma_t) s_i^{l'}, \tag{3}$$

where s_i^g is the logits from global branch, σ_t is a learnable scalar for fusion.

2 GPT-4V Prompt

The prompt used for GPT-4V to generate navigation reasoning, discussed in section §3.3 is shown in Figure 1.

[5] Refer to section 3.2 in the main paper.

{image}
As an AI navigating an indoor environment, you're given the task {instruction}. You find yourself at a particular juncture within the execution of this command. Based on your current observation of the surroundings, including obstacles, pathways, and relevant landmarks, determine the next step toward completing this task. Your response should briefly describe your immediate environment and specify the direction or action you will take to proceed. Summarize this in a concise paragraph, integrating both your observation and decision-making process.

Fig. 1: Navigation reasoning generation prompt for GPT-4V.

Table 1: Comparison of single-run performance on R2R dataset.

Methods	#	Val Seen					Val Unseen				
		TL	NE↓	OSR↑	SR↑	SPL↑	TL	NE↓	OSR↑	SR↑	SPL↑
w/o Visual-Language-Action Pretrain:											
DUET	1	12.38	3.62	73	66	60	13.20	4.07	72	64	55
w/o local branch	2	11.43	3.50	74	67	62	12.08	4.08	71	62	54
w/ EVA-CLIP-g	3	12.64	3.73	73	66	60	14.27	4.07	72	63	54
NavGPT-2$_{\text{FlanT5-XL}}$ (ours, 1.5B)	4	13.02	3.34	74	69	62	13.68	3.37	74	68	56
NavGPT-2$_{\text{FlanT5-XXL}}$ (ours, 5B)	5	13.08	2.98	79	74	65	13.25	3.18	80	71	60

3 Additional Rerults

In this section, we conduct additional experiments to illustrate the choice of navigation policy network for NavGPT-2 and the effectiveness of LLM features. To align the same training schema of the navigation policy, we conduct the experiments for DUET initiating it from LXMERT without VLN specialized pretraining.

3.1 Effect of Vision Encoder

Because NavGPT-2 exploits a stronger vision encoder [2], we conduct an ablation study on the original DUET to investigate the performance gain brought by the vision encoder. As shown in Table 1, after switching the visual representation to the stronger vision feature same as NavGPT-2, little performance gain is observed for the DUET global branch (Model # 3 compared to Model # 2). We hypothesize this is due to the global branch for DUET performing vision-language alignment on a coarse scale, while the fine-grained alignment is performed in the local branch. Therefore, the main performance gain in NavGPT-2 is not contributed by the stronger vision encoder but the better representation from LLM hidden.

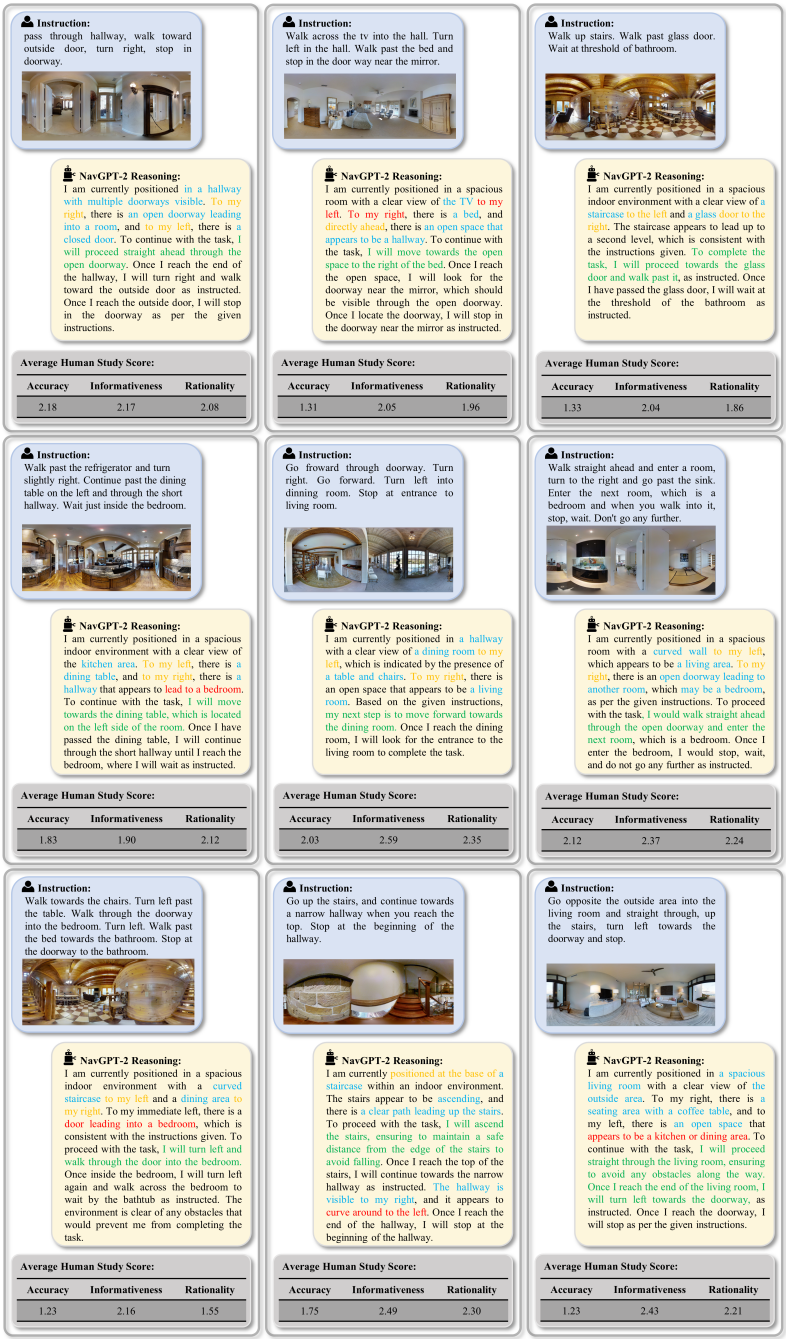

Fig. 2: Qualitive Results for NavGPT-2. It can correctly recognize object and scenes and their corresponding locations, grouding the observation to the given instruction and plan the next step. However, hallucination of the non-existent object or misjudged the direction is also observed.

3.2 Effect of VLN Pretrain

We consider the same training scale and the same training schema of DUET as NavGPT-2, without pertaining auxiliary VLN tasks and directly finetuning on the VLN dataset. Under the same training schema and scale of data, NavGPT-2 performs significantly better than the original DUET, shown in Table 1. This showcases the superiors of LLM features that enable the learning of cross-modality alignment in the downstream task when the visual feature is projected to the LLM's language space by the Q-former. Without VLN tailored pertaining, the performance of DUET significantly drops. We leave adding the pertaining process for the downstream navigation policy in future work.

3.3 Additional Qualitive Results

In this section, we present extra qualitative results in addition to §4.3. In Figure 2, we present the navigational reasoning produced by NavGPT-2 during navigation. NavGPT-2 is capable of forming a detailed understanding of its surroundings with objects and scenes and their corresponding orientations. Furthermore, it adeptly reasons about the progress of navigation and correlates it with specific portions of the instruction. Impressively, it is also able to predict expected observations, such as "appears to lead to a bedroom," based on the current visual inputs. This demonstrates NavGPT-2's ability not only to navigate but also to anticipate and interpret complex environments intelligently.

4 Limitations and Future Work

Although NavGPT-2 could generate navigation reasoning to some extent, it is hard to evaluate the effectiveness of these reasonings, since it is set as a single-step reasoning based on local observation and does not model the navigation history in the VLM. Instead, such history information is encoded in the downstream navigation policy. As a result, the consistency between navigation reasonings is underexplored. Moreover, the reasoning and action predicted by downstream navigation policy are not strictly synchronized in NavGPT-2, such synchronization could be done either explicitly by tuning LLM with the same supervision signal of action or by collaborating with the reasoning generation loss during fine-tuning the downstream policy network, we leave the synchronization to future work. Finally, the communicative capability of NavGPT-2 is not evaluated in this work, we suggest investigating the communicative ability of LM-based VLN agents and the synchronization between their reasoning and actions as a future direction.

5 Broader Effect

Our research endeavors to leverage Large Vision-Language Models (VLM) to develop VLN agents, while preserving the linguistic prowess of VLMs for explaining action predictions in natural language. We posit that the inherent communicative

capability, commonsense knowledge, and broad linguistic comprehension of VLM constitute the cornerstone for creating instruction-following navigation agents with generalizability. NavGPT-2 illuminates the reasonings of VLM throughout the navigation process explicitly and interpretably. Due to safety and ethical considerations, we currently conduct all experiments using the open-source Vision-and-Language Navigation dataset within a simulated environment, which ensures controlled agent behavior. Concurrently, we acknowledge that the potential practical application of this technology warrants further exploration, particularly in terms of action and reasoning synchronization, which remains an underexplored area. Notably, we observe the propensity of VLMs to hallucinate non-existent scenes or objects and fail to identify object directions, shown in Figure 2, which is also a common issue within VLM research. Future investigations are essential to address how to harmonize VLM action and reasoning and to enhance the agent's ability to self-explain in a manner intelligible to humans, a critical consideration for ensuring safety in real-world applications.

References

1. Chen, S., Guhur, P.L., Tapaswi, M., Schmid, C., Laptev, I.: Think global, act local: Dual-scale graph transformer for vision-and-language navigation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 16537–16547 (2022)
2. Fang, Y., Wang, W., Xie, B., Sun, Q., Wu, L., Wang, X., Huang, T., Wang, X., Cao, Y.: Eva: Exploring the limits of masked visual representation learning at scale. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 19358–19369 (2023)
3. Tan, H., Bansal, M.: Lxmert: Learning cross-modality encoder representations from transformers. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). pp. 5100–5111 (2019)