

AwareVLN: Reasoning with Self-awareness for Vision-Language Navigation

Supplementary Material

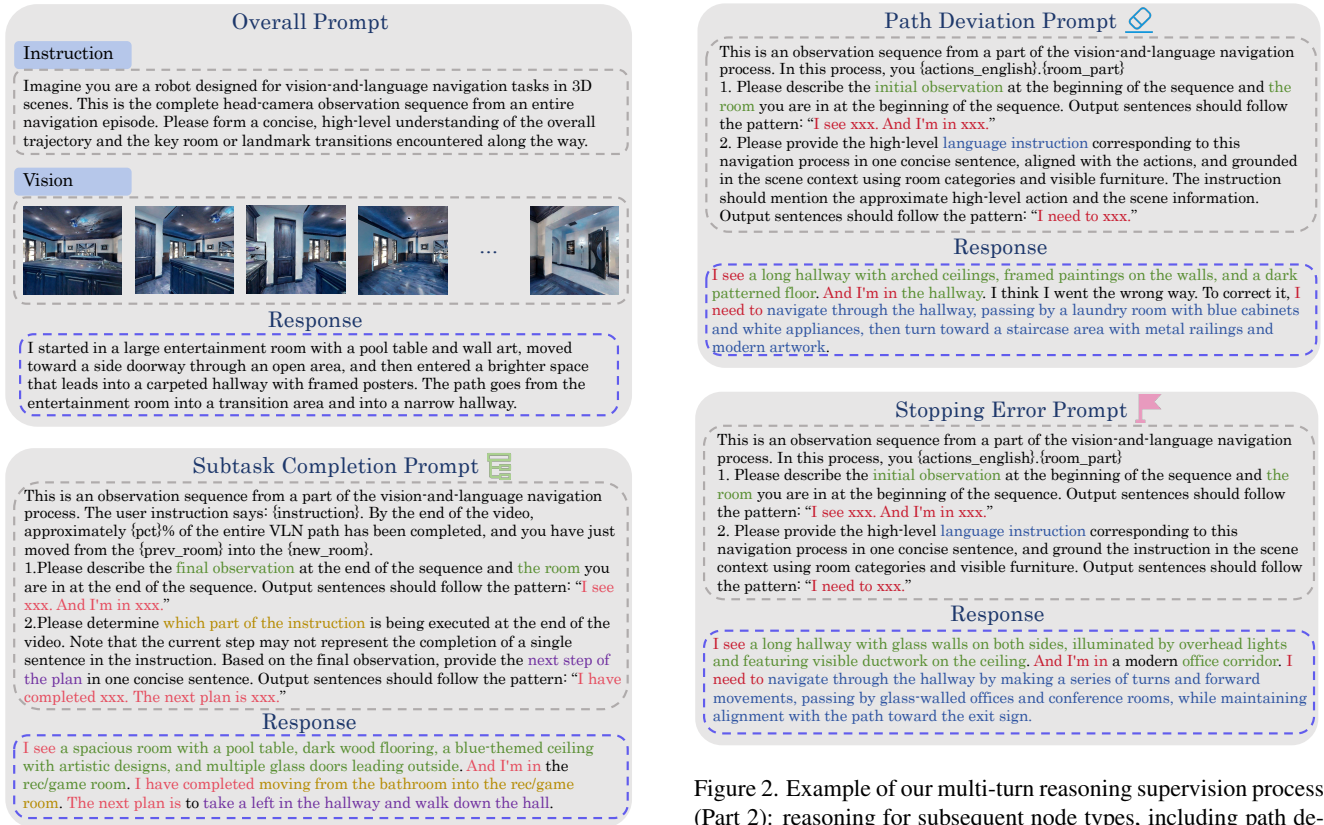


Figure 1. Example of our multi-turn reasoning supervision process (Part 1): global understanding of the navigation episode and reasoning for subtask completion based on localized observations.

A. Example of Multi-turn Reasoning Generation

To further illustrate the effectiveness of our data engine, we provide a concrete example of the multi-turn VLM interaction used for reasoning supervision generation, as shown in Figure 1 and Figure 2. Starting from the complete episode observation sequence, the VLM first produces a global summary of the navigation trajectory. Then, for each automatically detected key node, we prompt the model with localized visual inputs and node-specific context. The VLM subsequently generates structured reasoning outputs for different cases, including subtask completion, path deviation, and stopping error. This example demonstrates how our prompting design progressively guides the model to understand the navigation state, identify potential issues, and produce reasoning feedback aligned with real navigation progress.

Figure 2. Example of our multi-turn reasoning supervision process (Part 2): reasoning for subsequent node types, including path deviation and stopping error, demonstrating error interpretation and recovery planning.

B. Cross-Dataset Generalization

To further validate the generalization ability of different methods, we conduct a cross-dataset experiment in Table 1. All models are trained exclusively on RxR-CE training set and then evaluated on RxR-CE Val-Unseen split. Our AwareVLN also achieves leading performance under this transfer setting, demonstrating strong robustness.

C. Visualization of Automatically Collected Trajectories

Figure 3 and Figure 4 present two representative navigation trajectories collected by our automatic data engine. Each example follows a natural language instruction and illustrates the generated reasoning supervision at key navigation nodes. The visual observations, together with the VLM outputs, clearly demonstrate the agent’s understanding of the correction process, subtask completion, and next-step planning, validating the effectiveness of our data engine in pro-

Table 1. Cross-dataset performance on the RxR-CE Val-Unseen split. All results are obtained without training on RxR-CE.

	Observation			RxR Val-Unseen			
	S.RGB	Depth	Odo.	NE ↓	OS ↑	SR ↑	SPL ↑
LAW [5]	✓	✓	✓	10.87	21.0	8.0	8.0
CM2 [3]	✓	✓	✓	8.98	25.3	14.4	9.2
WS-MGMap [1]	✓	✓	✓	9.83	29.8	15.0	12.1
Seq2Seq [4]	✓	✓		11.8	5.02	3.51	3.43
CMA [4]	✓	✓		11.7	10.7	4.41	2.47
RGB-Seq2Seq [4]	✓			11.2	12.2	0.0	0.0
RGB-CMA [4]	✓			9.55	14.8	0.0	0.0
A ² NAV [2]	✓			-	-	16.8	6.3
NaVid [6]	✓			8.41	34.5	23.8	21.2
NaVILA	✓			8.78	46.8	34.3	28.2
AwareVLN (Ours)	✓			7.15	51.0	39.8	36.0

ducing scalable and interpretable supervision. Furthermore, the examples highlight how the agent analyzes navigation status. This showcases not only its ability to align reasoning with navigation progress but also the robustness of the data engine in capturing realistic behaviors and generating fine-grained supervision signals that benefit training self-aware navigation models.

References

- [1] Peihao Chen, Dongyu Ji, Kunyang Lin, Runhao Zeng, Thomas H Li, Mingkui Tan, and Chuang Gan. Weakly-supervised multi-granularity map learning for vision-and-language navigation. *arXiv preprint arXiv:2210.07506*, 2022. 2
- [2] Peihao Chen, Xinyu Sun, Hongyan Zhi, Runhao Zeng, Thomas H Li, Gaowen Liu, Mingkui Tan, and Chuang Gan. a² nav: Action-aware zero-shot robot navigation by exploiting vision-and-language ability of foundation models. *arXiv preprint arXiv:2308.07997*, 2023. 2
- [3] Georgios Georgakis, Karl Schmeckpeper, Karan Wanchoo, Soham Dan, Eleni Miltsakaki, Dan Roth, and Kostas Daniilidis. Cross-modal map learning for vision and language navigation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 15460–15470, 2022. 2
- [4] Jacob Krantz, Erik Wijmans, Arjun Majundar, Dhruv Batra, and Stefan Lee. Beyond the nav-graph: Vision and language navigation in continuous environments. In *European Conference on Computer Vision (ECCV)*, 2020. 2
- [5] Sonia Raychaudhuri, Saim Wani, Shivansh Patel, Unnat Jain, and Angel Chang. Language-aligned waypoint (law) supervision for vision-and-language navigation in continuous environments. In *Proceedings of the 2021 conference on empirical methods in natural language processing*, pages 4018–4028, 2021. 2
- [6] Jiazhao Zhang, Kunyu Wang, Rongtao Xu, Gengze Zhou, Yicong Hong, Xiaomeng Fang, Qi Wu, Zhizheng Zhang, and He Wang. Navid: Video-based vlm plans the next step for vision-and-language navigation. *Robotics: Science and Systems*, 2024. 2

Instruction: Walk down the hall passing the stairwell. Walk into the bedroom ahead that has a bed with white comforter and blue floor rug underneath.

I see a large window with a dark view outside, a purple abstract painting on the wall, and a doorway leading to a warmly lit room. And I'm in a hallway. I think I went the wrong way. To correct it, I need to turn left and move forward through the hallway toward the bedroom area.



I see a large abstract painting on the wall, a doorway leading into a bedroom with a bed covered in a white comforter, and a blue floor rug beneath it. And I'm in a hallway. The next plan is to enter the bedroom through the doorway ahead.



I see a spacious bedroom with a bed covered in a white comforter, a blue floor rug beneath it, a seating area with a sofa, and large windows showing a dark outdoor view. And I'm in a bedroom. I have completed walking down the hall passing the stairwell and entering the bedroom ahead that has a bed with a white comforter and blue floor rug underneath. The next plan is to explore the bedroom further.

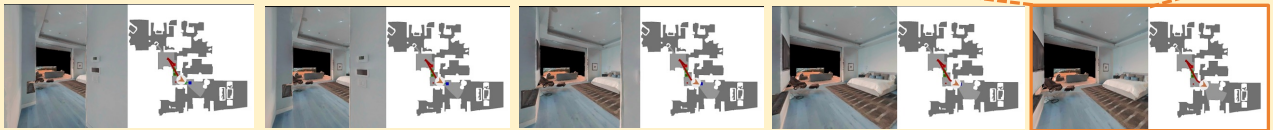


Figure 3. Example 1 of an automatically collected training trajectory, illustrating three key nodes: path deviation, correction completion, and subtask completion. The agent interprets the evolving visual scene and progressively generates structured reasoning outputs aligned with the navigation instruction.

Instruction: Enter the house and walk straight past the living room and staircase. When you get to the dining room table, turn left. Enter the room in front of you and stop in front of the table with the vases and open book on it.

I see a spacious living area with modern furniture, including sofas, a coffee table, and artwork on the walls, along with a staircase to the right and an open hallway extending forward. And I'm in the entryway/foyer/lobby. I have completed entering the house and walking straight past the living room and staircase. The next plan is to continue moving forward toward the dining room table.



I see a long, bright hallway with light-colored flooring, white walls, recessed lighting, and a modern staircase on the right side. There are abstract artworks on the walls and doors along the left side, leading to various rooms. And I'm in the hallway. I have completed entering the house and walking straight past the living room and staircase. The next plan is to continue moving forward toward the dining room table.



I see a spacious dining area with a long table, chairs, and large windows on the right side allowing natural light. There is artwork on the ceiling and walls, and the space opens into a larger room ahead. And I'm in the dining room. I have completed entering the house and walking straight past the living room and staircase. The next plan is to turn left at the dining room table.



I see a well-lit office space with a central table, an open book, vases, and decorative items on it. There is a large window with sheer curtains, a blue wardrobe on the left, and a comfortable seating area with a sofa and side table. And I'm in the office. I have completed entering the house, walking straight past the living room and staircase, turning left at the dining room table, and entering the room in front of me. The next plan is to stop in front of the table with the vases and open book on it.

Figure 4. Example 2 of an automatically collected training trajectory. This case includes multiple subtask completion nodes and room transitions, demonstrating the model's ability to reason over spatial changes and identify subtask boundaries during navigation.