

1 Appendix

1.1 Examples of Dynamic Chain-of-Navigation

The Figure 1, 4, 7 and 10 show how to leverage large language model GPT-4 to plan Dynamic Chain-of-Navigation (DCoN) based on the raw instruction and observed scene objects. Every input prompt in the above blue box comprises **Robot Definition**, **Navigation Strategy**, **Prediction Format**, and **Episode Information**. For benchmark evaluation, three groups of In-Context Learning Examples are given following the "User" and "Assistant" roles. The prediction is displayed in the bottom orange box, including reason, action, landmark, and flag. As defined in the prompt, the flag indicates whether the robot has completed the given instruction.

1.2 Examples of Multimodal Large Model's Prediction

The Figure 2, 5, 8 and 11 demonstrate how the multimodal large model GPT-4V predicts the next navigation direction based on the next action and landmarks from Dynamic Chain-of-Navigation and the panorama at the current position. Every multimodal prompt in the above blue box comprises panorama, task definition, navigation instruction, the next action and landmarks specified by the DCoN. The orange box at the bottom shows the prediction from the multimodal large model, which contains the thinking process and the next navigation direction.

1.3 Examples of Multi-sourced Value Maps

The Figure 3, 6, 9 and 12 displays examples of Multi-sourced Values Maps for different types of navigation instructions. These Multi-sourced Value Maps include the **Action Value Map**, **Semantic Value Map**, **Trajectory Value Map**, and **Intuition Value Map**. The decision-making value map at the current decision step is obtained by summing all these value maps. The position with the highest value on the decision-making value map is chosen as the next waypoint, which is annotated as a ★.

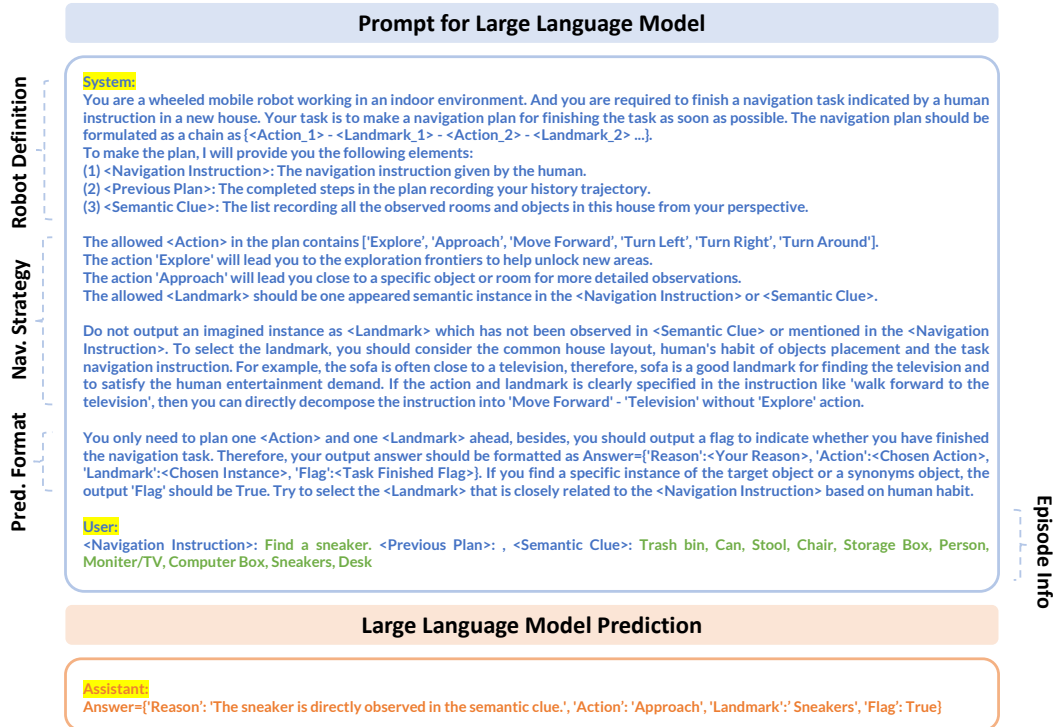


Figure 1: The case of Dynamic Chain-of-Navigation in the real world.

Multimodal Prompt for Multimodal Large Model

User:



You are an indoor navigation agent. I give you a panoramic observation image, complete navigation instruction and sub-instruction you should execute now. Direction 1 and 11 are ahead, Direction 5 and 7 are back, Direction 3 is to the right, and Direction 9 is to the left. Please carefully analyze visual information in each direction and judge which direction is most suitable for next movement according to the act and landmark mentioned in the sub-instruction.

Your answer should follow "Thinking Process" and "Judgement". In the "Judgement: " field, you should only write down direction ID you choose. If you think you have arrived the destination, you can answer "Stop" in the "Judgement: " field.

Note that the "Direction 5" and "Direction 7" are the directions you just came from. Generally, the direction with more navigation landmarks in the complete navigation instruction is better.

Given complete navigation instruction: "{Find a sneaker.}", action you should execute: "{Approach - Sneakers}".

Multimodal Large Model Prediction

Assistant:

Thinking Process: The complete navigation instruction requires finding a sneaker, and the action to be executed is "Approach - Sneaker". Among the directions provided, Direction 7 clearly shows a sneaker on the floor near a desk, which matches the target object mentioned in the navigation instruction.

Judgement: 7

Figure 2: The case of multimodal large model's prediction in the real world.

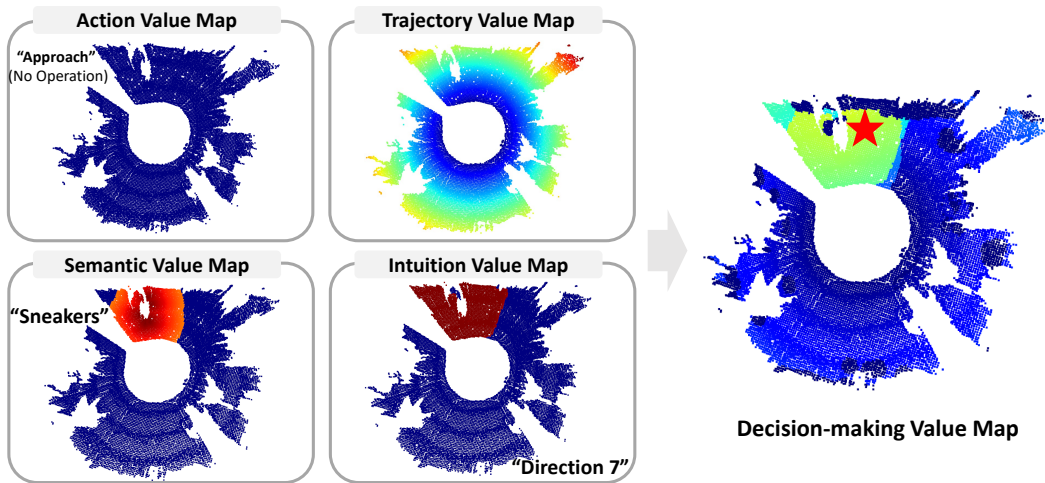


Figure 3: The case of Multi-sourced Value Maps in the real world based on Figure 1 and 2.



Figure 4: The case of Dynamic Chain-of-Navigation for object goal navigation.

Multimodal Prompt for Multimodal Large Model

User:



You are an indoor navigation agent. I give you a panoramic observation image, complete navigation instruction and sub-instruction you should execute now. Direction 1 and 11 are ahead, Direction 5 and 7 are back, Direction 3 is to the right, and Direction 9 is to the left. Please carefully analyze visual information in each direction and judge which direction is most suitable for next movement according to the act and landmark mentioned in the sub-instruction.

Your answer should follow "Thinking Process" and "Judgement". In the "Judgement:" field, you should only write down direction ID you choose. If you think you have arrived the destination, you can answer "Stop" in the "Judgement:" field.

Note that the "Direction 5" and "Direction 7" are the directions you just came from. Generally, the direction with more navigation landmarks in the complete navigation instruction is better.

Given complete navigation instruction: "{Find a toilet}", action you should execute: "{Explore -Bed}"

Multimodal Large Model Prediction

Assistant:

Thinking Process:

The sub-instruction is to "Explore - Bed." In the panoramic observation image, Direction 1 shows a living room setting with sofas but no visible bed. Direction 3 also shows part of the living room without a bed. Direction 5 presents a dining area leading into an open kitchen, still no bed in sight. Direction 7 offers a wider view of the dining and kitchen area, with no bed visible. Direction 9 shows a hallway with closed doors that could potentially lead to rooms with beds, but the beds themselves are not observable from this angle. Direction 11, on the other hand, shows a different angle of the hallway with one door slightly ajar, opening into a room which might contain a bed, but from this distance and angle, the contents of the room are unclear.

Judgement: 11

Figure 5: The case of multimodal large model's prediction for object goal navigation.

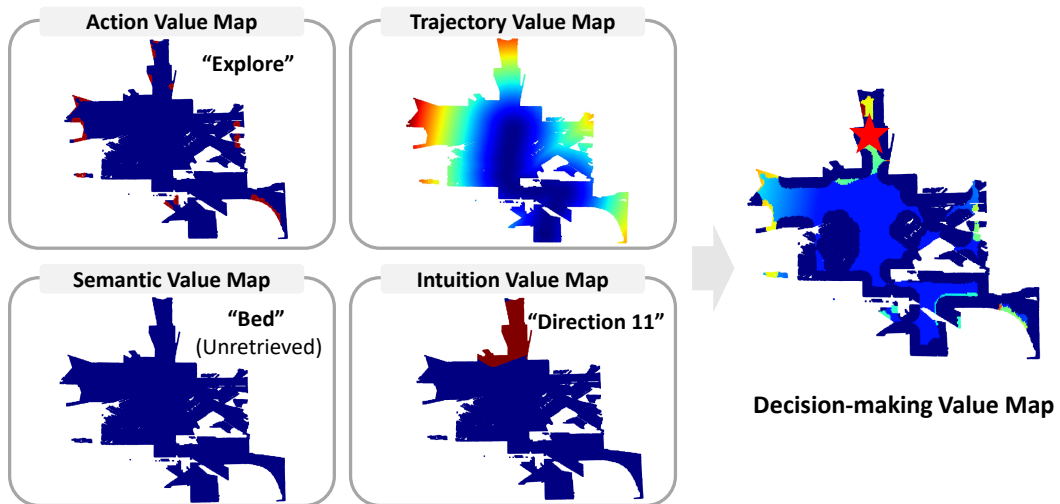


Figure 6: The case of Multi-sourced Value Maps for ObjNav based on Figure 4 and 5.



Figure 7: The case of Dynamic Chain-of-Navigation for visual language navigation.

Multimodal Prompt for Multimodal Large Model

User:



You are an indoor navigation agent. I give you a panoramic observation image, complete navigation instruction and sub-instruction you should execute now. Direction 1 and 11 are ahead, Direction 5 and 7 are back, Direction 3 is to the right, and Direction 9 is to the left. Please carefully analyze visual information in each direction and judge which direction is most suitable for next movement according to the act and landmark mentioned in the sub-instruction.

Your answer should follow "Thinking Process" and "Judgement". In the "Judgement: " field, you should only write down direction ID you choose. If you think you have arrived the destination, you can answer "Stop" in the "Judgement: " field.

Note that the "Direction 5" and "Direction 7" are the directions you just came from. Generally, the direction with more navigation landmarks in the complete navigation instruction is better.

Given complete navigation instruction: "[Go around the stairs into the hallway. Walk down the hallway until you reach the end, take a right, and stop at the entrance to the living room.]", action you should execute: "[Approach - Hallway]",

Multimodal Large Model Prediction

Assistant:

Thinking Process:

- Based on the navigation instruction to "walk down the hallway," we are looking for a direction that continues along the hallway.
- Direction 1 shows the continuation of the hallway with a turn at the end, which is consistent with the instruction to reach the end of the hallway and then take a right.
- Directions 3 and 9 show walls, indicating these are not the directions to walk down the hallway.
- Direction 5 and Direction 7 are to be avoided as they are directions from where we came.
- Direction 11 might show the continuation of the hallway, but it's visually similar to Direction 1, suggesting it's the same direction but the opposite side of the hallway.
- The correct direction should be the one that allows us to continue straight down the hallway without deviating through other spaces.

Judgement: Direction 1

Figure 8: The case of multimodal large model's prediction for visual language navigation.

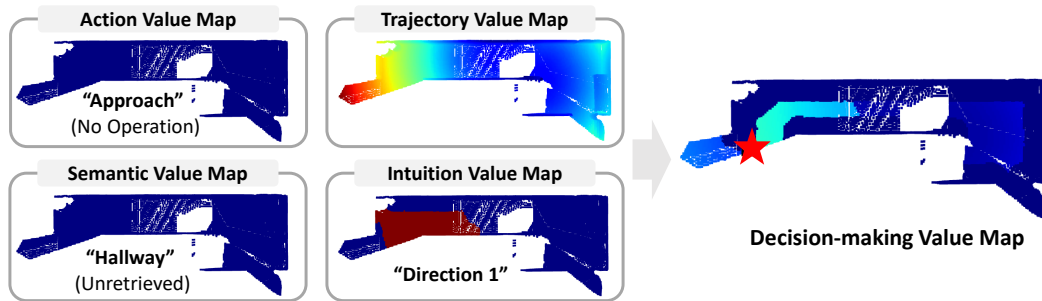


Figure 9: The case of Multi-sourced Value Maps for VLN based on Figure 7 and 8.



Figure 10: The case of Dynamic Chain-of-Navigation for demand-driven navigation.

Multimodal Prompt for Multimodal Large Model

User:



You are an indoor navigation agent. I give you a panoramic observation image, complete navigation instruction and sub-instruction you should execute now. Direction 1 and 11 are ahead, Direction 5 and 7 are back, Direction 3 is to the right, and Direction 9 is to the left. Please carefully analyze visual information in each direction and judge which direction is most suitable for next movement according to the act and landmark mentioned in the sub-instruction.

Your answer should follow "Thinking Process" and "Judgement". In the "Judgement: " field, you should only write down direction ID you choose. If you think you have arrived the destination, you can answer "Stop" in the "Judgement: " field.

Note that the "Direction 5" and "Direction 7" are the directions you just came from. Generally, the direction with more navigation landmarks in the complete navigation instruction is better.

Given complete navigation instruction: "[I wish to expand my knowledge.]", action you should execute: "[Approach - TV]".

Multimodal Large Model Prediction

Assistant:

Thinking Process: To satisfy the demand of expanding knowledge, an appropriate choice could be a place with books, educational material, or technology such as a computer and tv. Examining the provided panoramic observation, most directions show tables, chairs, and artwork on the walls, which do not directly suggest areas for expanding knowledge. However, Direction 3 displays a scene with an open door revealing a room that could possibly contain books, educational materials, or other resources for learning.

Judgement: Direction 3

Figure 11: The case of multimodal large model's prediction for demand-driven navigation.

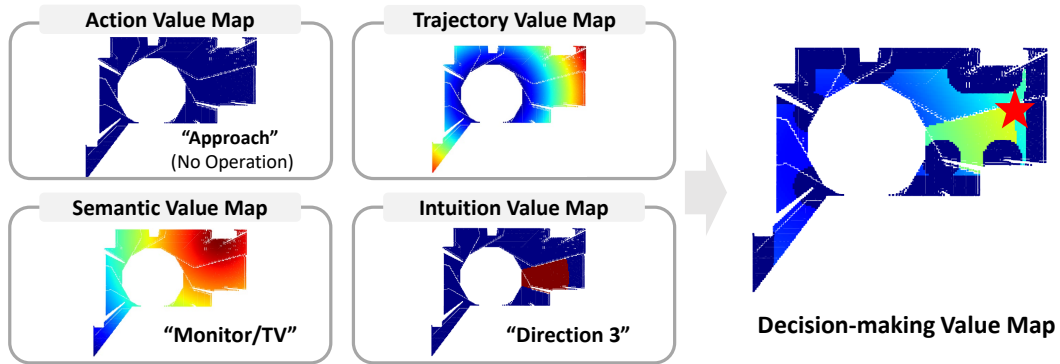


Figure 12: The case of Multi-sourced Value Maps for DDN based on Figure 10 and 11.