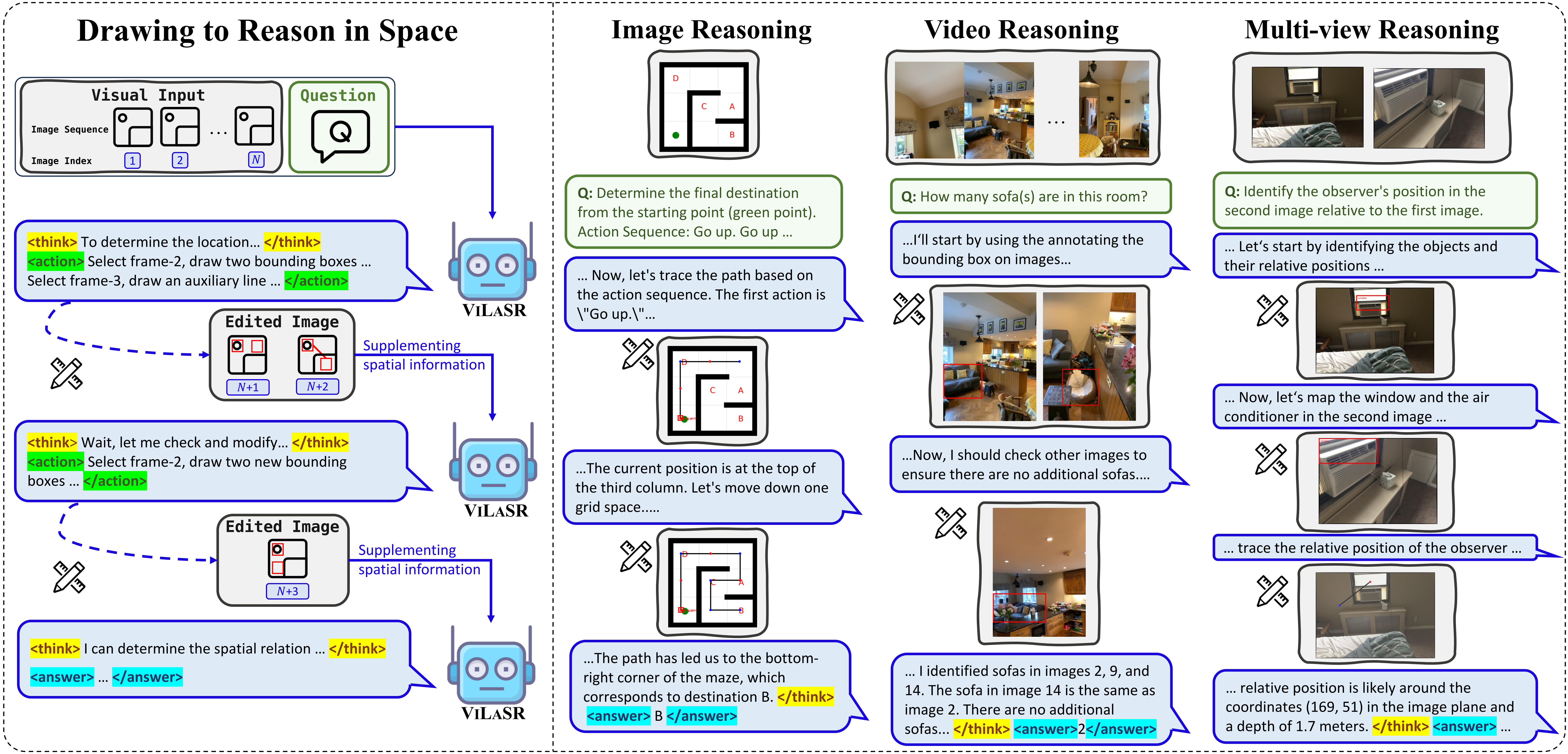

<think> To determine the location... </think> <action> Select frame-2, draw two bounding boxes ... Select frame-3, draw an auxiliary line ... </action>

<think> Wait, let me check and modify... </think> <action> Select frame-2, draw two new bounding boxes ... </action>

<think> I can determine the spatial relation ... </think> <answer> ... </answer>

Q: Determine the final destination from the starting point (green point). Action Sequence: Go up. Go up ...

... Now, let's trace the path based on the action sequence. The first action is \"Go up.\"...

...The current position is at the top of the third column. Let's move down one grid space.....

...The path has led us to the bottom-right corner of the maze, which corresponds to destination B. </think> <answer> B </answer>

Q: How many sofa(s) are in this room?

...I'll start by using the annotating the bounding box on images...

...Now, I should check other images to ensure there are no additional sofas....

... I identified sofas in images 2, 9, and 14. The sofa in image 14 is the same as image 2. There are no additional sofas... </think> <answer>2</answer>

Q: Identify the observer's position in the second image relative to the first image.

... Let's start by identifying the objects and their relative positions ...

... Now, let's map the window and the air conditioner in the second image ...

... trace the relative position of the observer ...

... relative position is likely around the coordinates (169, 51) in the image plane and a depth of 1.7 meters. </think> <answer> ...

Results from GPT-4o with only textual reasoning, which fails to preserve spatial details and dynamics *(highlighted in red words with error description)*:

Image Reasoning: *The final destination from the starting point is cell C*. Here is the step-by-step movement based on the given action sequence: Start at the green point. Go Up: Move to cell D. Go Up: Move to the wall (cannot move further but stay in cell D). *Go Right: Move to cell C. Go Right: Move to cell A. Go Down: Move to cell B. Go Left: Move back to cell A. Go Down: Move back to cell C. Go Right: Attempt to move to cell A again. The final position after following this sequence of moves is cell C. (Fails to accurately track position changes)*

Video Reasoning: There are two sofas in the room. *One is visible in image 2, and another can be seen in image 9. (Misses temporal-spatial correlations across frames)*

Multi-view Reasoning: … In the first image, we see the bed … from a more frontal view … In the second image, the angle is side-oriented and much closer to the air conditioner … *suggesting the observer moved to the right* and depth toward the air conditioner … *(Only captures coarse spatial relationships and fails to correctly reason about perspective transformations)*