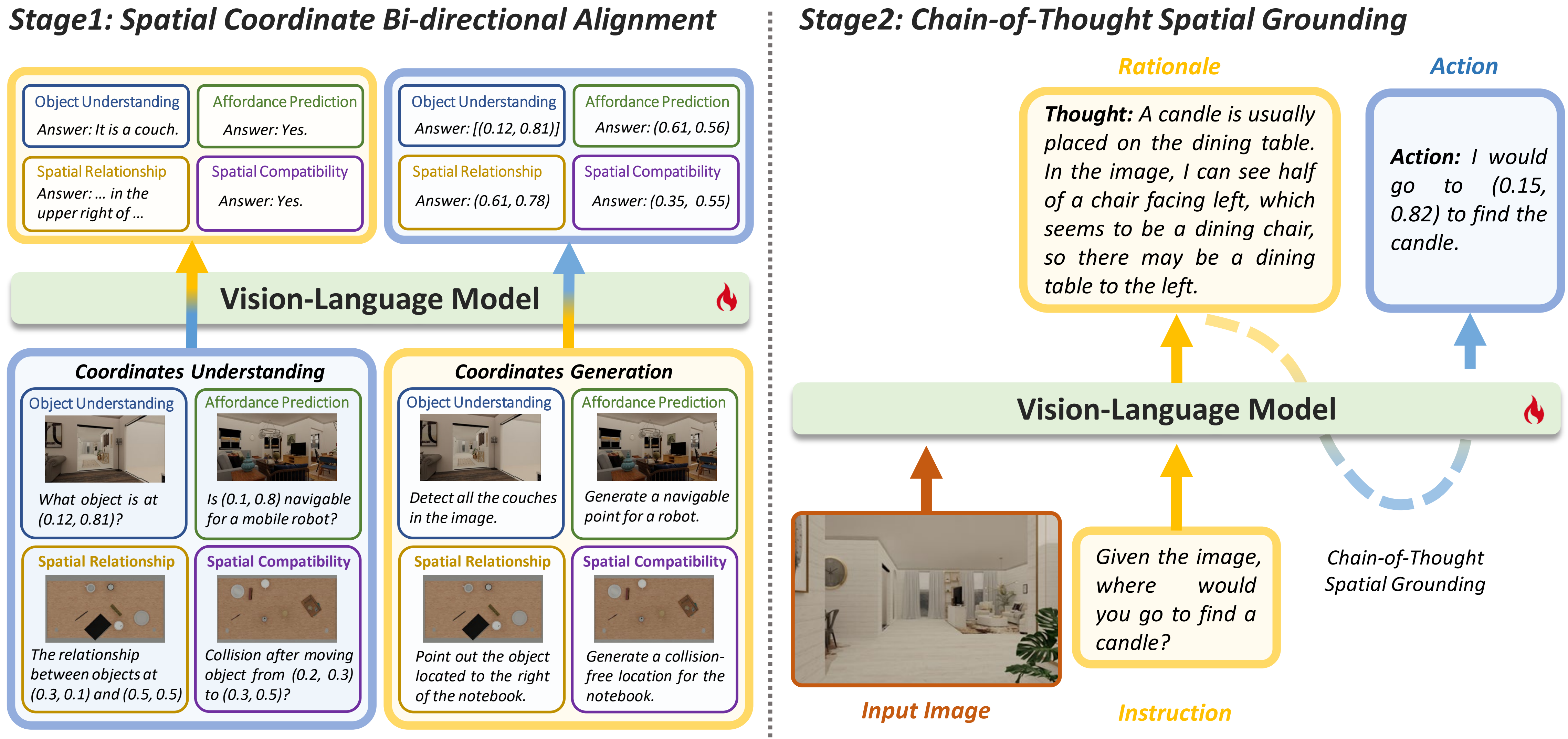
Stage1: Spatial Coordinate Bi-directional Alignment
Object Understanding
Answer: It is a couch.
Affordance Prediction
Answer: Yes.
Spatial Relationship
Answer: … in the upper right of …
Spatial Compatibility
Answer: Yes.
Object Understanding
Answer: [(0.12, 0.81)]
Affordance Prediction
Answer: (0.61, 0.56)
Spatial Relationship
Answer: (0.61, 0.78)
Spatial Compatibility
Answer: (0.35, 0.55)
Vision-Language Model
Coordinates Understanding
Object Understanding
What object is at (0.12, 0.81)?
Affordance Prediction
Is (0.1, 0.8) navigable for a mobile robot?
Spatial Relationship
The relationship between objects at (0.3, 0.1) and (0.5, 0.5)
Spatial Compatibility
Collision after moving object from (0.2, 0.3) to (0.3, 0.5)?
Coordinates Generation
Object Understanding
Detect all the couches in the image.
Affordance Prediction
Generate a navigable point for a robot.
Spatial Relationship
Point out the object located to the right of the notebook.
Spatial Compatibility
Generate a collision-free location for the notebook.
Stage2: Chain-of-Thought Spatial Grounding
Rationale
Action
Thought: A candle is usually placed on the dining table. In the image, I can see half of a chair facing left, which seems to be a dining chair, so there may be a dining table to the left.
Action: I would go to (0.15, 0.82) to find the candle.
Vision-Language Model
Given the image, where would you go to find a candle?
Chain-of-Thought Spatial Grounding
Input Image
Instruction

a) Spatial Coordinate Bi-directional Alignment

b) Chain-of-Thought Spatial Grounding