SpatialClaw

Oral Knowledge Questions

Explain each answer in your own words.

1. **What is an action interface?**

   Explain what an action interface determines about how a VLM invokes tools and receives their results.

   How does SpatialClaw combine the flexibility of writing Python with the ability to inspect results and revise its approach?

2. **How do segmentation and reconstruction work together?**

   Suppose the question is "How far is the chair from the table?"

   (a) What information does a segmentation mask provide, and what information does a 3D reconstruction provide?

   (b) How can the mask select the object's points from a dense 3D point map?

   (c) Why must the mask and point map correspond to the same frame, with measurements expressed in compatible coordinates?

3. **Why must the agent inspect intermediate results?**

   A tool returns a mask, and the next Python cell successfully computes a distance from it.

   (a) Why does this not establish that the answer is correct?

   (b) What can displaying the mask with *show()* reveal that printing only the distance cannot?

   (c) If the mask is wrong, how can the agent revise its analysis while reusing valid results already stored in the kernel?

4. **What does "training-free" mean for SpatialClaw?**

   SpatialClaw improves a VLM's spatial reasoning without updating its model weights.

   (a) Where does the additional capability come from?

   (b) Why does "training-free" not imply low inference cost?

   (c) How does SpatialClaw's use of explicit code and intermediate arrays differ from learning continuous visual thought tokens, as in CoVT?

5. **Why can a tool-using agent still give a wrong answer?**

   Suppose SpatialClaw returns an incorrect distance even though every tool call and Python cell completed successfully.

   (a) Give one possible **perception error** and one possible **reasoning or measurement error** that could explain the result.

   (b) What checks could help distinguish these two causes?

   (c) Why does access to accurate numerical computation not guarantee accurate spatial reasoning?