



QA-RP Pretraining

1) Local Geometry

Find the point covered by the **red cross**. Output the point's **3D coordinates**.

$[-0.9, -1.2, 2.7]$

2) Global Scene Context

Detect the **3D center points of objects** across all frames under the first camera coordinate system.

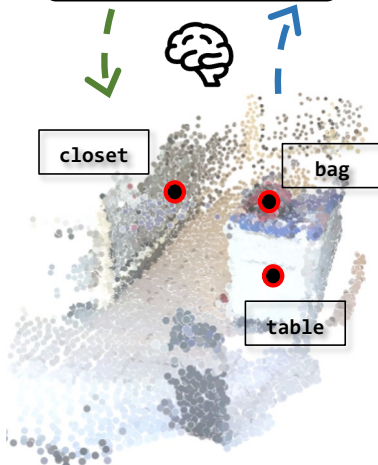
closet: $[1.5, -2.1, 4.3]$;
bag: $[0.4, -0.6, 1.8]$; . . .

Stage I

MLLM

Stage II

CoT-VC Finetuning



Activated Metric 3D Structure
in MLLM

what is the distance between the **closet** and the **bag** (in meters)?

closet: center = $(1.5, -2.1, 4.3)$ m,
dimensions = $3.4 \times 0.5 \times 2.1$ m

bag: center = $(0.4, -0.6, 1.8)$ m,
dimensions = $0.3 \times 0.2 \times 0.4$ m

Center distance: 3.1 m
Subtracting approximate
radii (1.0 m, 0.2 m):
Distance $\approx \max(0, 3.1 - 1.0 - 0.2)$
 $= 1.9$ m

<reliability: HIGH>

Final answer: 1.9