


## Input 3D Scene



 **Query:** "a brown leather chair located under the first dark brown table immediately inside the door ..."

## 3D Holistic View Selection



Multi-view  
Observation



0

...



n



: "In **view 1**, the layout of the room is visible, showing the tables and chairs most clearly ..."

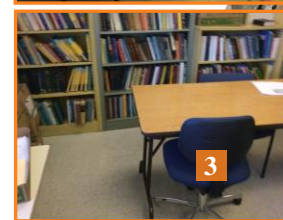
## 3D-2D Joint Decision-Making



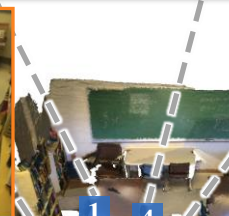
Camera Parameters:  $\mathbf{K}$ ,  $\mathbf{T}_{cw}$  ...



1



3

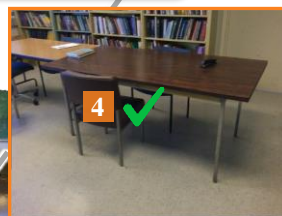


1

3

4

5



4



5



: "In **global view**, the chairs labeled 1, 4 are directly under this table ... Examining **camera images**, the chair labeled 4 is confirmed to be brown and leather ... Other chairs like 5 and 1, 3 do not match the color and material. So the best matching one is **chair 4**."

## Candidate Object Screening



Detected Objects



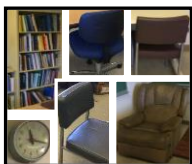
: "Target  
class: **chair**"

Matching

Anchor Filtering

if match **false**

Visual Object Table



1

3

4

5

7

8

9

→



1

3

4

5



: "The first dark brown table inside the door appears to be near the left of the image. The chairs under this table are labeled with IDs **1, 3, 4, and 5**..."