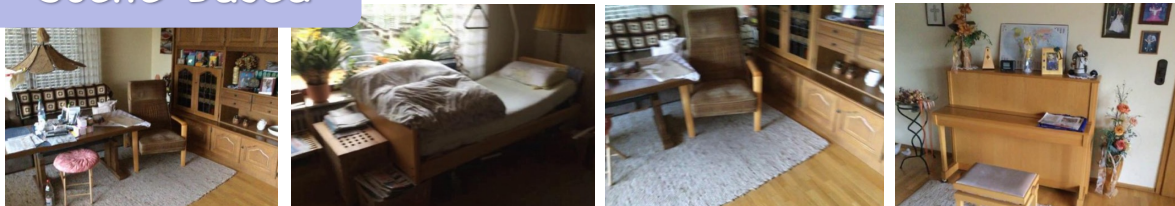



Static Scene

Scene-Based

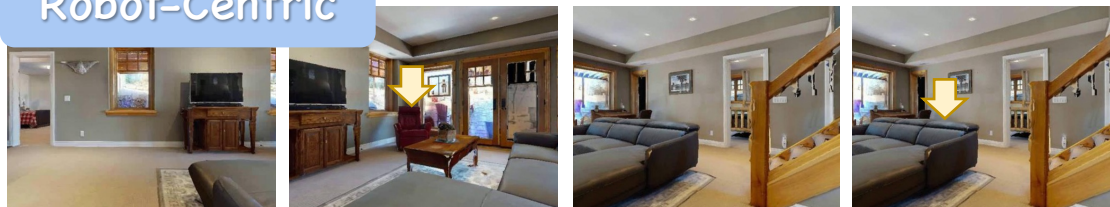


<Question>: What is the **main material** of the furniture in this room?

<Answer>: The main material of the furniture in this room overall is predominantly **wood**. 

Scene State

Robot-Centric

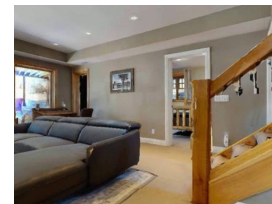


<Question>: **Which is farther from you**, the red sofa or the black sofa in the living room?

<Answer>: The **red sofa**. 

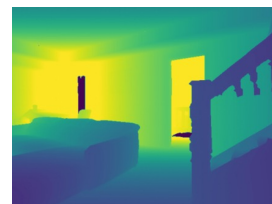
Distance Awareness

Input



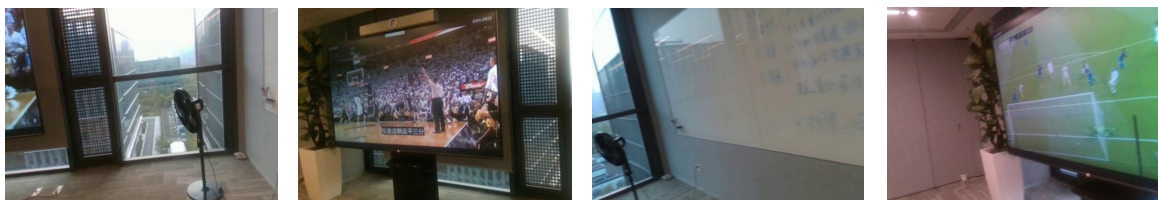
RGB

+



Depth

Dynamic Scene

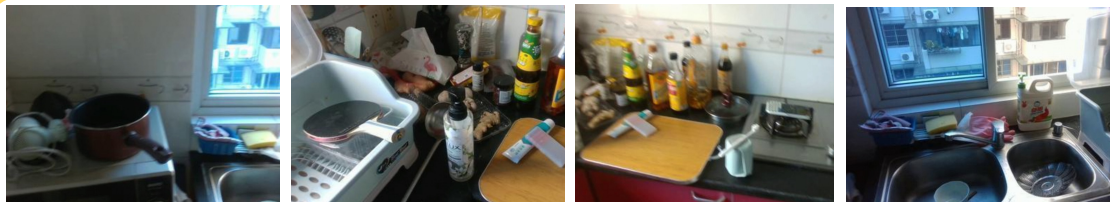


<Question>: **Before** you faced the whiteboard, what was the theme of the content **being played on the screen**?


<Answer>: **NBA basketball game**. 

Information Change

Hallucination



<Question>: Are there any **ping pong paddles** in this **kitchen**?

<Answer>: Yes.   **GPT-4o**: No. 

<Question>: Is the **toothbrush** next to the **stove** or on the **sink**?

<Answer>: Next to the stove.  Co-occurrence Countersense



Multilevel Eval.



Binary Eval.

