Multimodal Understanding

Visual Grounding

Q: Provide the bounding box coordinate of the police vehicle.
A: [0.26, 0.56, 0.44, 0.71]

Image Captioning

Q: Provide a one-sentence caption for the image.
A: A vintage-style street clock stands prominently at a city intersection, with a historic brick building in the background and several cars, including a police car, navigating the crosswalk.

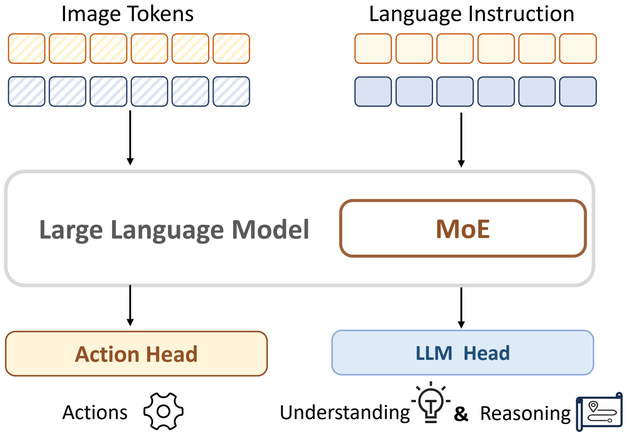
OCR

Q: What number is the hour hand pointed at?
A: The number 2 is pointed at.

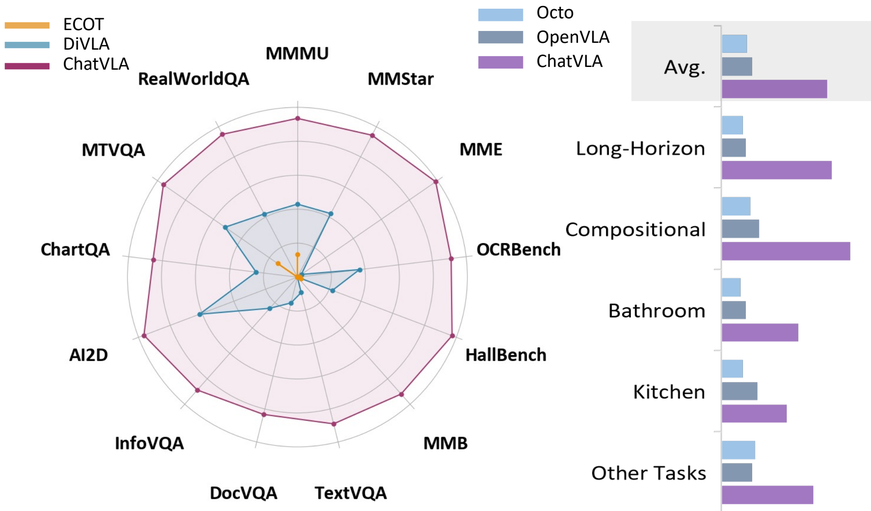
VQA

Q: What is the colors of the door in the building?
A: The door in the image is red.

ChatVLA Model



Results on Control & Understanding Dataset

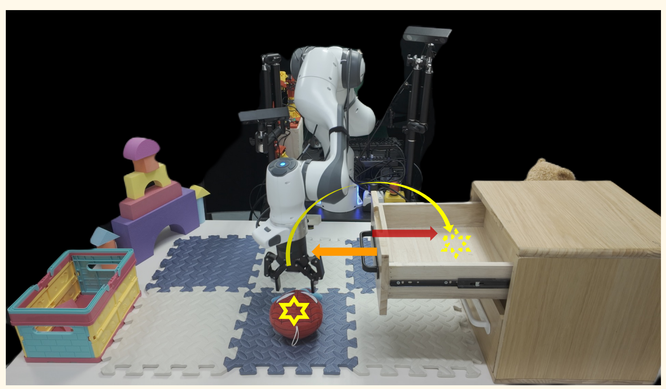


25 Real Robot Tasks

Long-Horizon Task with Direct Prompting

"Put the spider-man into the drawer."

1. *Open the drawer*
2. *Put the toy into it*
3. *Close the drawer*



Cross-Skill Multi-Tasking

Pick	Hang	Move	Stack	...
Bread	Block	Toy	Cup	Soap	Towel	...

Long-Horizon Task with High-Level Planner

Get the plate and place it on the tablecloth.
Flip the cup and place it on the tablecloth.
Move the bread to the plate.