MODEL CAPABILITY

EGOCENTRIC UNDERSTANDING
Perceive from user

SPATIO-TEMPORAL GROUNDING
Locate in space & time

PHYSICAL WORLD REASONING
Physical understanding

FINE-GRAINED ACTION PLANNING
Plan detailed steps

MULTIMODAL OUTPUT

REGION
Bounding Box
Segmentation Mask

TRAJECTORY
Motion Path
Pointing Sequence

POINTING
Area Prediction
Affordance Prediction

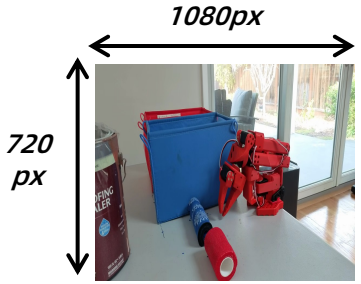
TEXT
Understanding
Reasoning

Dense / MoE Decoder

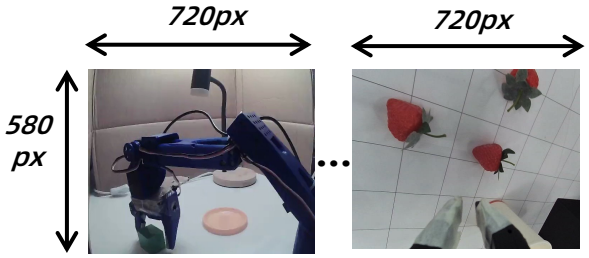
Vision Encoder

Tokenizer

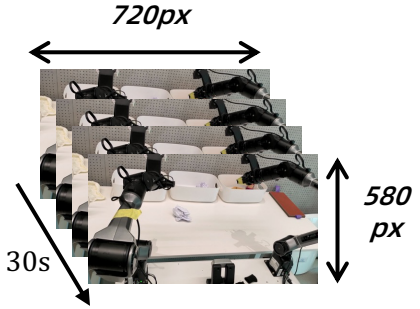
Omni Vision Input



Single-view Image



Multi-view Image



Video

Instructions

Q1: What action am I doing?

Q2: Where is the table at 10s?

Q3: Move the box to the sofa.

Q4: How to get to the kitchen?