



Question x_t :

Taking into account all the actions performed by c, what can you deduce about the primary objective and focus within the video content?

Video x_v :



Candidate answer set Y :

- y_1 : C is cooking
- y_2 : C is doing laundry
- y_3 : C is cleaning the kitchen
- y_4 : C is cleaning dishes
- y_5 : C is cleaning the bathroom

MVU

Frame Selection

Per Frame Object Detection

hand located at
(0.39, 0.7, 0.02)
dish located at
(0.55, 0.62, 0.096) ...

Object Matching

hand trajectory:
(0.39,0.7,0.02)-> ...
dish trajectory:
(0.55,0.62,0.096)-> ...

Template Operations

- Question and Candidates x_t, Y
- Most Relevant Frame x_v^l**
- Global Object Information x_{GOI}**
- Object Spatial Location x_{OSL}**
- Object Motion Trajectory x_{OMT}**

VLM Answer Selection

e_i

y_1 y_2 y_3 y_4 y_5

e_1 e_2 e_3 e_4 e_5

\hat{y} : C is cleaning dishes