



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?

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

Batch

s_1	x_t	y_1
s_2	x_t	y_2
s_3	x_t	y_3
s_4	x_t	y_4
s_5	x_t	y_5

LLM / VLM
Single Pass w/ Mask

Predicted Logits

p_1	
p_2	
p_3	
p_4	
p_5	

Cross Entropy Error only on y_i

x_t	y_1	e_i
x_t	y_2	
x_t	y_3	
x_t	y_4	
x_t	y_5	

$\operatorname{argmin}_{y_i \in Y} \text{CE}(p_i, y_i)$

y_1	e_1
y_2	e_2
y_3	e_3
y_4	e_4
y_5	e_5



C is cleaning dishes