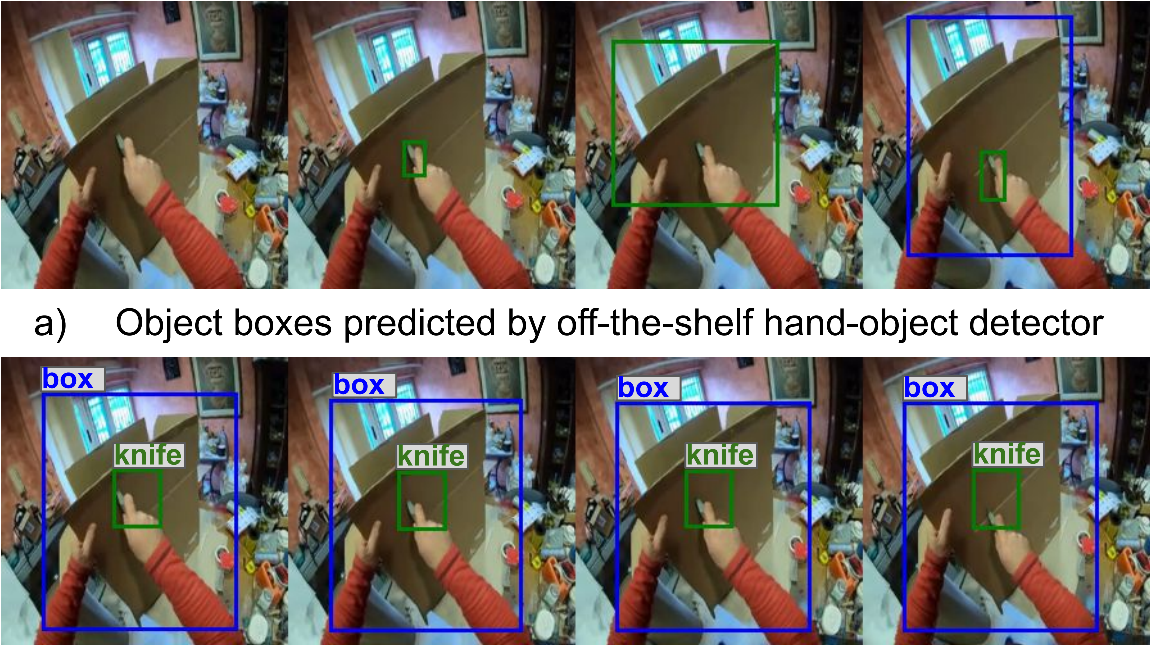
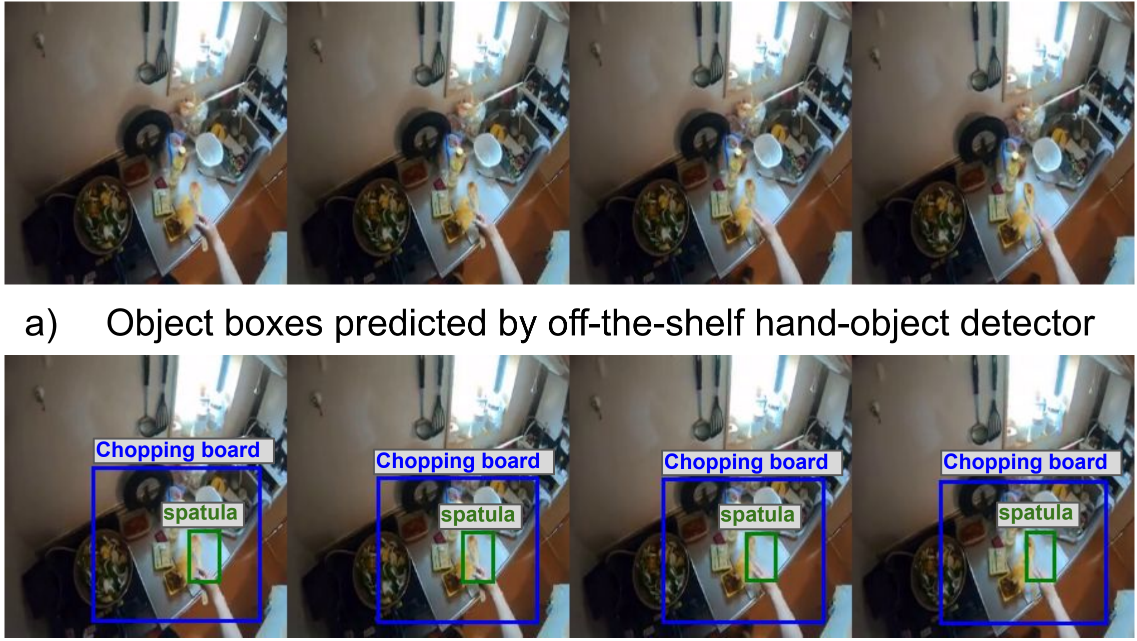
Text: #C C cuts the box with the utility knife



a) Object boxes predicted by off-the-shelf hand-object detector

b) Vision-text grounding results from the proposed model

#C C puts the spatula in his right hand on the chopping board



a) Object boxes predicted by off-the-shelf hand-object detector

b) Vision-text grounding results from the proposed model