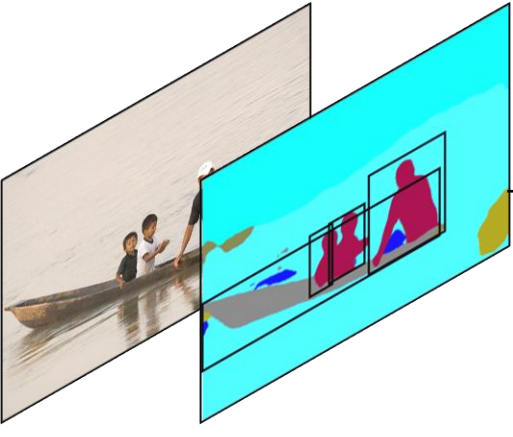
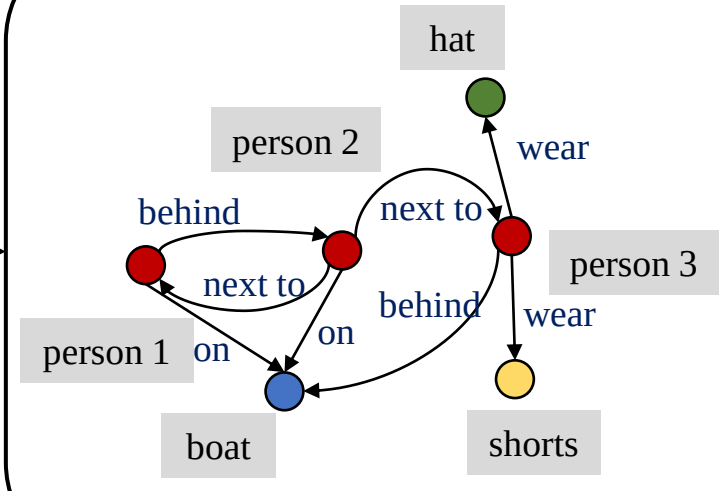


Q1: How many persons are there ?
A1: Three.
Q2: What is the man wearing ?
A2: A hat.

……

Q10: Where are they?
A10: On a boat.



Original image

Image input w/ missing vison

Dialog interaction as supplementary

Generated scene graph