



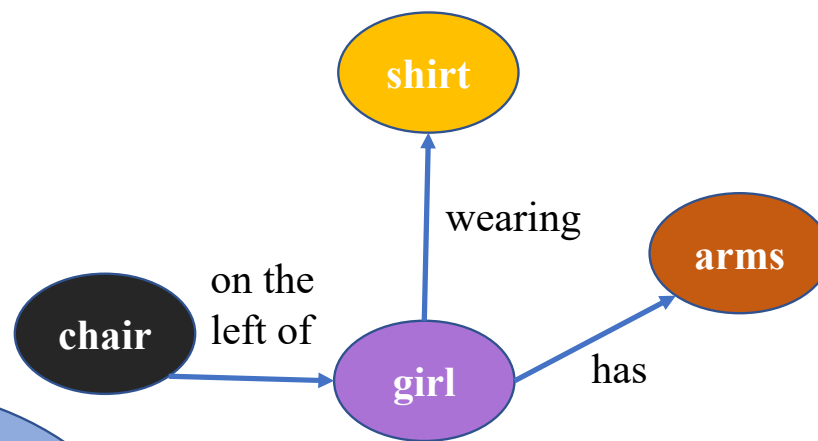
Image

+

“what does the
image describe
about girl, shirt,
arms, chair?”

Text

Scene Graph
Generation
Model



**girl is wearing shirt,
is has arms. Chair is
on the left of girl.**

Example result