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

Image + Text

Scene Graph Generation Model

Example result

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