Although existing image caption models can produce promising results using recurrent neural networks (RNNs), it is difficult to guarantee that an object we care