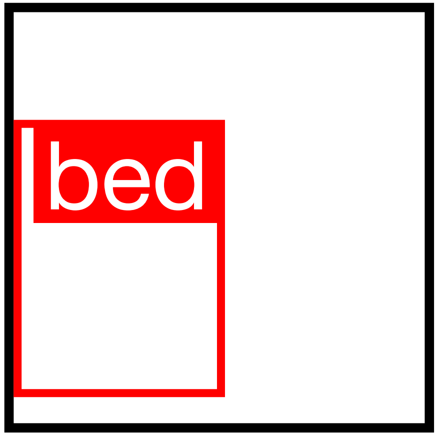
Caption input:
*A room with **two beds***



Weakly condition on masked RGB and strongly on caption & bounding box