Current image captioning systems perform at a merely descriptive level, essentially enumerating the objects in the scene and their relations. Humans, on the co