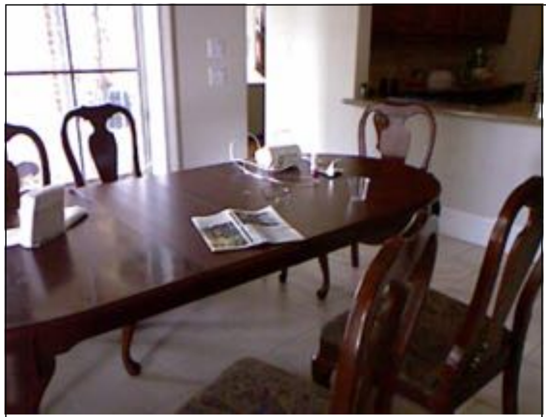
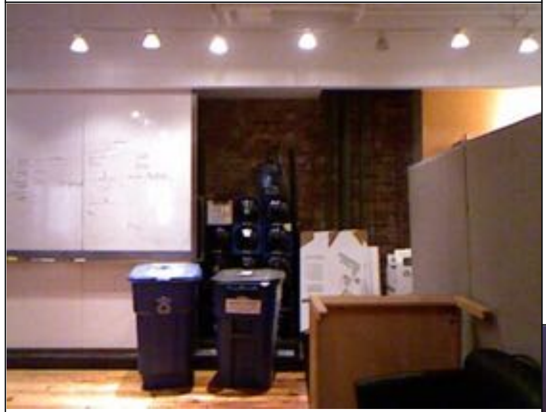
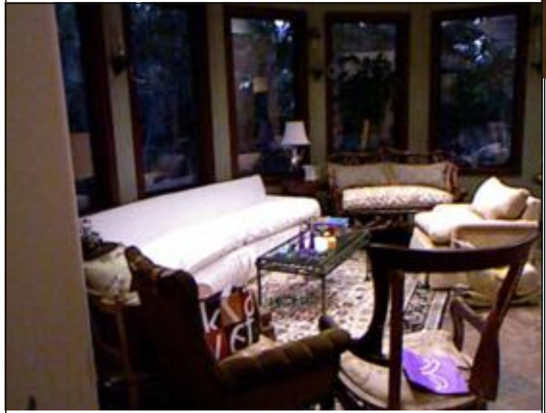


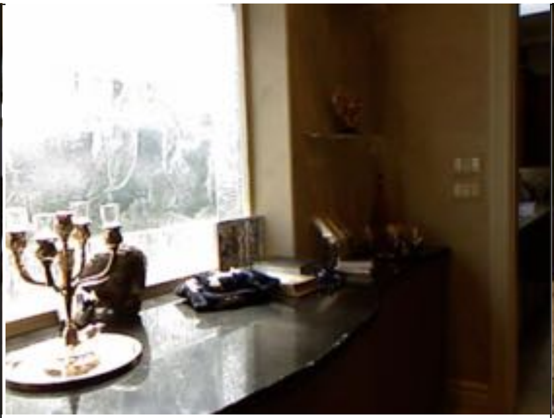
QA: (What is behind the table?, window)
Spatial relation like 'behind' are dependent on the reference frame. Here the annotator uses observer-centric view.



QA: (what is behind the table?, sofa)
Spatial relations exhibit different reference frames. Some annotations use observer-centric, others object-centric view
QA: (how many lights are on?, 6)
Moreover, some questions require detection of states 'light on or off'

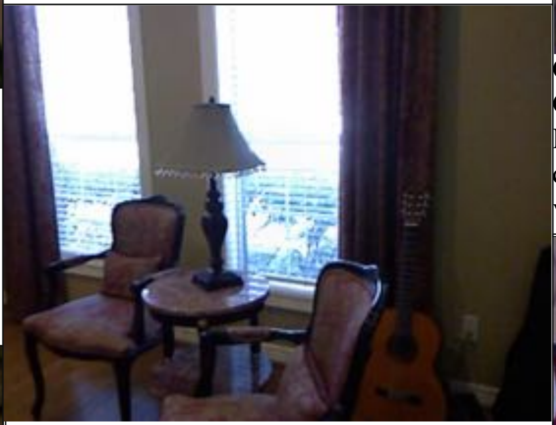


Q: what is at the back side of the sofas?
Annotators use wide range spatial relations, such as 'backside' which is object-centric.



QA: (what is beneath the candle holder, decorative plate)
Some annotators use variations on spatial relations that are similar, e.g. 'beneath' is closely related to 'below'.

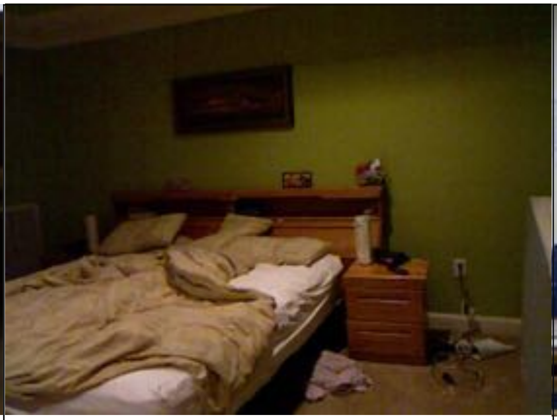
QA: (what is in front of the wall divider?, cabinet)
Annotators use additional properties to clarify object references (i.e. wall divider). Moreover, the perspective plays an important role in these spatial relations interpretations.



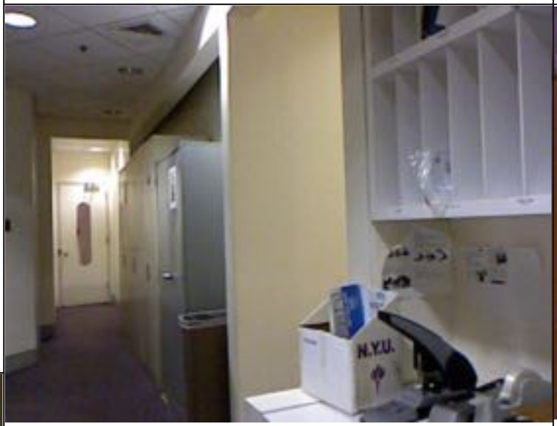
QA1: (what is in front of the curtain behind the armchair?, guitar)

QA2: (what is in front of the curtain?, guitar)

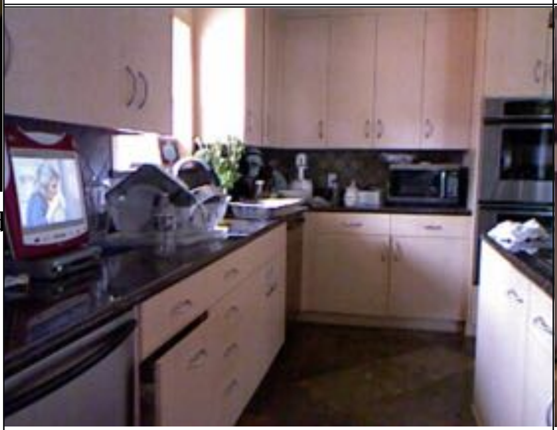
Spatial relations matter more in complex environments where reference resolution becomes more relevant. In cluttered scenes, pragmatism starts playing a more important role



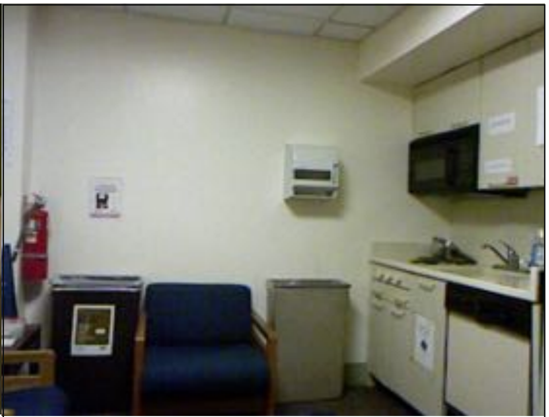
The annotators are using different names to call the same things. The names of the brown object near the bed include 'night stand', 'stool', and 'cabinet'.



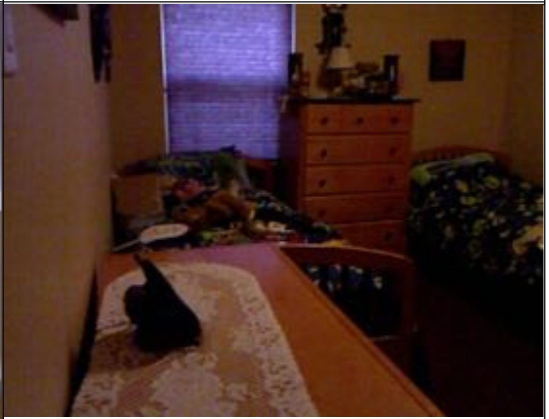
QA1:(How many doors are in the image?, 1)
QA2:(How many doors are in the image?, 5)
Different interpretation of 'door' results in different counts: 1 door at the end of the hall vs. 5 doors including lockers



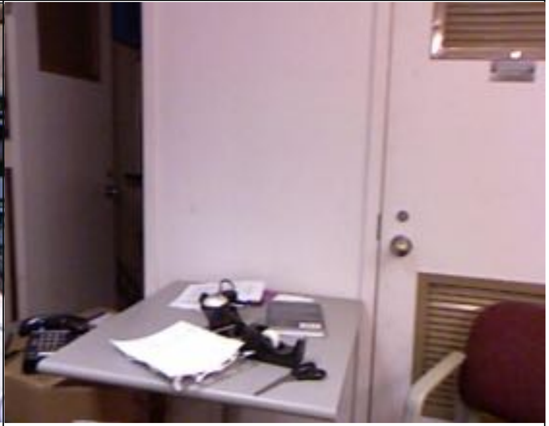
QA: (What is the object on the counter in the corner?, microwave)
References like 'corner' are difficult to resolve given current computer vision models. Yet such scene features are frequently used by humans.



Some objects, like the table on the left of image, are severely occluded or truncated. Yet, the annotators refer to them in the questions.



QA: (How many drawers are there?, 8)
The annotators use their common-sense knowledge for amodal completion. Here the annotator infers the 8th drawer from the context



QA: (How many doors are open?, 1)
Notion of states of object (like open) is not well captured by current vision techniques. Annotators use such attributes frequently for disambiguation.