VLM Hallucinations

Oral Knowledge Questions

Explain each answer in your own words.

1. **What counts as a VLM hallucination?**

   Explain the difference between an unsupported object claim, an incorrect attribute, and an incorrect relation. Give one example of each.

   Why can a fluent, plausible description still be unreliable?

2. **Why can familiar expectations override an image?**

   Suppose a VLM describes a visibly black banana as yellow. Explain how learned associations could contribute to this error.

   How does Visual Contrastive Decoding use predictions from the original and a distorted image to reduce this kind of mistake? Does the method require retraining the VLM?

3. **Where can visual grounding fail?**

   Distinguish between these two situations:

   - The visual encoder does not adequately preserve a detail needed to answer the question.
   - The detail is available in the visual representation, but the language model fails to use it appropriately.

   Why might improving the visual encoder help the first situation but fail to resolve the second?

4. **Testing whether a VLM responds to visual changes**

   Two images show a pair of lines. In the original image, the lines have equal length. In the edited image, the lower line is longer. The model answers "Yes" to "Are the lines the same length?" for both images.

   (a) Why does the correct answer on the original image provide insufficient evidence that the model compared the line lengths?

   (b) What does the incorrect answer on the edited image reveal? Does this result alone prove that the model ignored the image entirely?

   (c) Why does requiring correct answers on both images provide a stronger test than giving credit for either answer separately?

5. **Distinguishing truthfulness from helpfulness**

An image clearly shows four dogs, but their breed cannot be determined reliably. The user asks, "How many dogs are shown?"

Consider these responses:

- **A:** "There are dogs."
- **B:** "There are four dogs."
- **C:** "There are four golden retrievers."

(a) Which response contains supported information but fails to provide the information requested?

(b) Which response directly answers the question without adding an unsupported claim?

(c) Why should Response C receive less credit for truthfulness than Response B, even though both give the correct count?