

the results from second run are largely consistent with the observations from the first run, reinforcing the main conclusions about the differing behaviors of BLIP-2 and LLaVA in these specific conflict scenarios.

Here's a breakdown of the consistency:

1. **Text Conflict:**
   * **BLIP-2:** Highly consistent.
     + Agreement with Falsehood: 1.4% (Run 1) vs 1.5% (Run 2) - Still very low.
     + Confusion/Irrelevance: 72.9% (Run 1) vs 80.0% (Run 2) - Remains the dominant behavior.
     + Correct Rejection: 25.7% (Run 1) vs 18.5% (Run 2) - Still a minority behavior.
     + *Conclusion:* BLIP-2 consistently tends towards irrelevant or confused responses when faced with this type of textual conflict, rarely agreeing with the falsehood.
   * **LLaVA:** Broadly consistent pattern, but with a noticeable shift.
     + Agreement with Falsehood: 29.2% (Run 1) vs 47.7% (Run 2) - Increased significantly.
     + Confusion/Irrelevance: 0.0% (Run 1) vs 0.0% (Run 2) - Consistently avoids this category.
     + Correct Rejection: 48.3% (Run 1) vs 52.3% (Run 2) - Remained relatively stable (and slightly increased).
     + Processing Error: 22.5% (Run 1) vs Not explicitly shown/likely 0% (Run 2) - This difference is important. If the errors from Run 1 were resolved, the outcomes for those samples in Run 2 could explain the shift towards higher agreement.
     + *Conclusion:* LLaVA consistently engages directly (either agreeing or rejecting), avoiding confusion. However, the tendency to *agree* with the falsehood became more pronounced in the second run, potentially influenced by the samples that previously caused errors. The core finding that it directly addresses the prompt, unlike BLIP-2, remains consistent.
2. **Image Conflict:**
   * **BLIP-2:** Highly consistent.
     + Ignored Perturbation: 61.0% (Run 1) vs 58.0% (Run 2) - Very similar majority behavior.
     + Other/Irrelevant Description: 39.0% (Run 1) vs 42.0% (Run 2) - Very similar minority behavior.
     + *Conclusion:* BLIP-2 consistently tends to ignore the specific image perturbations (rotation + blur) used in the experiment, prioritizing the original semantic content.
   * **LLaVA:** Highly consistent.
     + Ignored Perturbation: 18.0% (Run 1) vs 14.0% (Run 2) - Consistently low.
     + Other/Irrelevant Description: 82.0% (Run 1) vs 86.0% (Run 2) - Consistently the dominant behavior.
     + *Conclusion:* LLaVA consistently shows sensitivity to the image perturbations, rarely ignoring them, but its descriptions often fall into the 'Other/Irrelevant' category based on the evaluation heuristic.

**Overall Consistency:**

The core behavioral differences identified in the first run are confirmed in the second run:

* BLIP-2 is evasive in text conflicts and robust/ignorant to these image perturbations.
* LLaVA is direct (though sometimes incorrect) in text conflicts and sensitive to image perturbations (though its descriptions of perturbed images are often categorized as 'Other').

The main point of slight inconsistency is the increased rate at which LLaVA agreed with the textual falsehood in the second run, which might be linked to resolving processing errors encountered in the first run or simple variance in the random sample selection if it differed slightly. However, the fundamental patterns of behavior for both models remain consistent across the two runs.