Overall Assessment:

These new results are significantly better in terms of providing a more reliable and nuanced understanding of the models' behaviors. The larger, category-balanced sample size (400 vs. initial 10 or 100) reduces the impact of random chance or bias from a small selection of images. The refined evaluation categories also capture model responses more accurately. The trends observed now are much more likely to represent the general tendencies of BLIP-2 and LLaVA on this type of task.

Detailed Analysis and Comparison:

1. Text Conflict (Misleading Captions):

BLIP-2:

Consistency: Still extremely resistant to explicitly agreeing with the falsehood (0.6%, similar to 1.4-1.5% before).

Change: The tendency towards "Confusion/Irrelevance" became even more pronounced, rising to 90.3% (from ~73-80%). Correspondingly, "Correct Rejection" decreased further to 9.1% (from ~18-26%).

Interpretation: On a larger, more diverse dataset, BLIP-2's strategy of avoiding direct engagement with the conflicting prompt is even clearer. It overwhelmingly defaults to generating unrelated output rather than judging the misleading statement. Its architecture, particularly the Q-Former acting as a bridge, might filter or process the conflicting inputs in a way that leads to this evasive behavior rather than a direct comparison.

LLaVA:

Consistency: Still engages directly with the prompt, largely avoiding "Confusion/Irrelevance" (0.3%). The rate of "Correct Rejection" remains substantial and relatively stable (42.6% vs. ~48-52% before).

Change: "Agreement with Falsehood" remains very high at 52.6% (similar to 47.7% in Run 2, higher than 29.2% in Run 1). The new "Implicit Rejection" category captures a small fraction (4.6%) of responses.

Interpretation: LLaVA consistently attempts to answer the verification question. However, it appears highly susceptible to being swayed by the textual prompt, agreeing with the falsehood more often than rejecting it correctly. This could reflect a text-dominance issue, potentially exacerbated by visual instruction tuning which might prioritize following textual instructions. The simple MLP projector in LLaVA might pass through conflicting signals more directly than BLIP-2's Q-Former.

2. Image Conflict (Perturbed Images - Rotation + Blur):

BLIP-2:

Consistency: Still fails to explicitly acknowledge the perturbations using the defined keywords (0.0% "Acknowledged Perturbation").

Change: A significant shift occurred. "Ignored Perturbation" dropped considerably to 29.9% (from ~58-61%). Consequently, "Other/Irrelevant Description" became the dominant category at 70.1% (up from ~39-42%).

Interpretation: This is a key refinement. While BLIP-2 rarely describes the image exactly as if unperturbed anymore, it still doesn't describe the perturbation itself. Instead of ignoring the changes, it now more often generates descriptions that are simply different from the original and don't mention the rotation/blur. This suggests the perturbations do affect its processing, but not in a way that leads it to identify and articulate the nature of the visual change. It might be losing confidence or defaulting to more generic descriptions.

LLaVA:

Consistency: Still rarely ignores the perturbation (4.1%, similar to ~14-18% before). "Other/Irrelevant Description" remains the most frequent outcome (75.7%, similar to ~82-86%).

Change: The most striking finding is the emergence of a significant "Acknowledged Perturbation" rate (20.2%).

Interpretation: This confirms LLaVA is sensitive to the visual changes and, in a notable minority of cases (~1 in 5), it can and does describe the perturbation using terms like "rotated" or "blurred". This capability was likely hidden in the broader "Other/Irrelevant" category in the previous, less nuanced analysis. Its visual instruction tuning might contribute to this ability to describe visual characteristics when prompted. However, the fact that "Other/Irrelevant" is still the largest category indicates that even when acknowledging the change isn't happening, its description often deviates significantly or fails the heuristic check.

Refined Conclusions (Based on Run 3):

Text Conflict: The larger dataset confirms BLIP-2's strong tendency towards evasive, irrelevant responses, while LLaVA directly engages but is highly prone (over 50% of the time) to agreeing with the textual falsehood, suggesting a potential text-dominance issue.

Image Conflict: The larger dataset reveals a more nuanced picture. BLIP-2 is affected by perturbations more than initially thought, but responds with irrelevant descriptions rather than ignoring them or describing the change. LLaVA confirms its sensitivity and demonstrates a clear capability (in ~20% of cases) to explicitly identify the type of global perturbation applied, a behavior not seen in BLIP-2.

In summary, these latest results are much more informative due to the improved experimental setup. They solidify the core differences in how BLIP-2 and LLaVA handle these conflicts, providing clearer evidence for BLIP-2's evasiveness/irrelevance and LLaVA's direct engagement (prone to text influence) and its partial ability to describe visual perturbations.

Latest Run Result:

latest run result

--- Text Conflict Results Summary ---

Agreement with Falsehood (Count) Agreement with Falsehood (%) \

model

BLIP-2 2 0.6

LLaVA 173 52.6

Confusion/Irrelevance (Count) Confusion/Irrelevance (%) \

model

BLIP-2 297 90.3

LLaVA 1 0.3

Correct Rejection (Count) Correct Rejection (%) \

model

BLIP-2 30 9.1

LLaVA 140 42.6

Implicit Rejection (Count) Implicit Rejection (%)

model

BLIP-2 0 0.0

LLaVA 15 4.6

--- Image Conflict Results Summary ---

Acknowledged Perturbation (Count) Acknowledged Perturbation (%) \

model

BLIP-2 0 0.0

LLaVA 79 20.2

Ignored Perturbation (Count) Ignored Perturbation (%) \

model

BLIP-2 117 29.9

LLaVA 16 4.1

Other/Irrelevant Description (Count) Other/Irrelevant Description (%)

model

BLIP-2 274 70.1

LLaVA 296 75.7