Based on the results you provided from running the experiment, here's a conclusion comparing BLIP-2 and LLaVA on handling modality conflicts in this specific setup:

--- Text Conflict Results Summary ---

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

model

BLIP-2 1 1.4

LLaVA 26 29.2

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

model

BLIP-2 51 72.9

LLaVA 0 0.0

Correct Rejection (Count) Correct Rejection (%) \

model

BLIP-2 18 25.7

LLaVA 43 48.3

Processing Error (Count) Processing Error (%)

model

BLIP-2 0 0.0

LLaVA 20 22.5

--- Image Conflict Results Summary ---

Ignored Perturbation (Count) Ignored Perturbation (%) \

model

BLIP-2 61 61.0

LLaVA 18 18.0

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

model

BLIP-2 39 39.0

LLaVA 82 82.0

**1. Text Conflict (Misleading Captions):**

* **LLaVA's Direct Approach:** LLaVA appears to engage more directly with the verification task posed by the prompt ("Is this [misleading statement]?"). It achieved a higher rate of "Correct Rejection" (48.3%) compared to BLIP-2 (25.7%), suggesting it was more often successful in identifying the textual falsehood based on the visual evidence. However, when LLaVA failed, it was significantly more likely to exhibit "Agreement with Falsehood" (29.2%) than BLIP-2 (1.4%). This indicates that while LLaVA attempts the task more faithfully, it's also more susceptible to being explicitly misled by the text when it doesn't correctly prioritize the visual information.
* **BLIP-2's Evasive Behavior:** BLIP-2 showed a much lower rate of agreeing with the falsehood. However, its dominant failure mode was "Confusion/Irrelevance" (72.9%). This suggests that when faced with this specific type of textual conflict and prompt, BLIP-2 often defaulted to generating unrelated or nonsensical output rather than directly agreeing or disagreeing based on the conflict.
* **LLaVA's Processing Errors:** A notable portion of LLaVA samples resulted in "Processing Error" (22.5%). This could indicate issues with the model handling certain inputs, potential instability with the quantization level used, or problems specific to the prompt format for those cases within the Colab environment. This reduces the effective number of samples compared for LLaVA.

**Conclusion (Text Conflict):** LLaVA attempts to answer the verification question more directly, leading to more correct rejections but also more instances of explicitly agreeing with the false text. BLIP-2 is less likely to agree with the falsehood but often avoids directly addressing the conflict, resulting in irrelevant output.

**2. Image Conflict (Perturbed Images):**

* **BLIP-2's Semantic Persistence:** BLIP-2 demonstrated a strong tendency to "Ignore Perturbation" (61.0%). This means it frequently described the scene based on its original semantic content, effectively overlooking the applied visual distortions (rotation and blur). This suggests a stronger reliance on learned priors or less sensitivity to these specific visual changes. The remaining 39.0% fell into "Other/Irrelevant Description," indicating outputs that didn't match the original scene nor (based on the heuristic) explicitly mention the perturbation.
* **LLaVA's Sensitivity to Perturbation:** LLaVA was much less likely to ignore the perturbation (18.0%). However, the vast majority of its responses (82.0%) were categorized as "Other/Irrelevant Description." This implies that while LLaVA *was* affected by the visual changes, it often failed to produce a description that either resembled the original scene *or* explicitly mentioned the perturbation keywords the evaluation checked for ("rotated," "blurred," etc.). It might be describing the visually altered scene in some way, but not in a manner that aligns closely with the original caption's meaning or uses the specific expected keywords for the perturbation type.

**Conclusion (Image Conflict):** BLIP-2 often prioritizes the learned semantic understanding of the scene over the perturbed visual input. LLaVA is more sensitive to the visual changes, but its descriptions of the perturbed images frequently deviate significantly from the original scene's description without necessarily identifying the specific nature of the perturbation as defined by the evaluation keywords.

**Overall Summary:**

This experiment suggests distinct modality handling strategies between the two models under these specific conflict conditions:

* **LLaVA:** Appears more sensitive to both textual instructions (leading to direct attempts at verification) and visual perturbations. However, this sensitivity can lead to more explicit errors (agreeing with falsehoods, generating descriptions irrelevant to the original scene after perturbation) and higher processing instability in this setup.
* **BLIP-2:** Seems more robust in generating *some* output and often defaults to a semantically "expected" description, showing less sensitivity to the specific textual conflict prompt (often becoming irrelevant) and the visual perturbations used (often ignoring them).

It's important to remember these conclusions are based on the specific conflicts generated (limited map, rotation/blur), the sample size, the heuristic evaluation methods, and the potential impact of quantization within the Colab environment. Different types of conflicts or evaluation metrics might reveal different aspects of the models' behavior.