Robustness of Vision-Language Models Under Cross-Modal Conflicts[‡]

Sreevidya Bollineni
sreevidyabol@umass.edu

Arin Garg
aringarg@umass.edu

Rati Rastogi
ratirastogi@umass.edu

Abstract

Vision-Language Models (VLMs) such as BLIP-2 and LLaVA are increasingly used for tasks that require joint reasoning over images and text. However, their reliability when presented with conflicting visual and textual information, termed cross-modal conflict, remains underexplored. This project investigates these vulnerabilities using a systematic evaluation on a balanced subset of COCO-2017 validation dataset, introduces a novel inference-time mitigation strategy leveraging a separate large language model, and demonstrates significant improvements in robustness. The findings highlight both the risks and the potential for practical solutions in real-world deployment.

1. Introduction

1.1. Problem Definition and Motivation

VLMs have achieved impressive results in image captioning, visual question answering and multimodal reasoning. Yet, their behavior under cross-modal conflict - when image and text inputs contradict each other - poses critical questions for reliability, especially in high-stakes applications. For instance, a model might be shown a photo of a dog but asked, "Is this a cat?" or presented with an image degraded by noise and asked for a description. In such scenarios, the model's ability to reconcile conflicting information is essential for trustworthy deployment

1.2. Research Questions

Our study addresses the following key questions:

1. How do VLMs respond to textual prompts that contradict the image content?
2. Do VLMs acknowledge visual perturbations in images or do they ignore them in favor of describing the semantic content?
3. Can we develop strategies to mitigate these vulnerabilities and improve model robustness?

[*]GitHub Repository: https://github.com/Sreevidya-B/RoubstnessOfVLMSUnderCrossModalConflicts

[†]Project Presentation: https://youtu.be/uRb4LegzlEs

1.3. Approach Overview

Our approach involves two primary experiments:

1. **Text Conflict Experiment:** We present models with images from the COCO dataset along with misleading textual questions that contradict the visual content. For example, showing an image of a dog and asking, "Is this a cat?", We then categorize and analyze the models' responses.
2. **Image Conflict Experiment:** We apply visual perturbations (primarily diffusion noise) to the images and ask models to describe them. We assess whether models acknowledge the perturbations or ignore them to describe the semantic content.

For both experiments, we evaluate two state-of-the-art VLMs: BLIP-2 and LLaVA. After establishing baseline performance, we develop and evaluate a mitigation strategy that leverages a separate large language model to generate targeted questions that help the VLMs recover accurate understanding.

1.4. Contributions

The primary contributions of this work include:

- A systematic evaluation of VLM robustness under cross-modal conflicts using a diverse set of COCO images covering all object categories.
- Quantification of vulnerability patterns in current state-of-the-art models.
- A novel mitigation strategy that significantly improves model robustness without requiring model retraining.
- Insights into the factors that influence model susceptibility to cross-modal conflicts.

2. Related work

Prior research has highlighted several aspects of VLM robustness and cross-modal reasoning:

- **Modality Prioritization:** Frank et al. (2021) [1] observed that VLMs often rely more on visual input for text understanding, but this integration is not always reliable, especially when modalities disagree.
- **Cross-Modality Conflicts:** Zhu et al. (2024) and others [2] introduced the concept of cross-modality knowledge

conflicts, showing that VLMs can be misled by textual claims and proposed dynamic contrastive decoding as a mitigation.

- **Commonsense Conflicts:** Liu et al. (2024) [3] found that VLMs ignore obvious visual clues because they rely too much on what they have learned from text. They developed Focus-on-Vision (FoV) prompting, a technique that helps models pay more attention to images when making decisions.
- **Perceptual Failures:** Anis et al. (2025) [4] tested whether VLMs understand basic image changes, such as rotation and brightness adjustments and found that they often fail to recognize these modifications. This weakness could make them less reliable in real-world applications where images are not always perfect.

How Our Work is Different:

Our work systematically evaluates both text-to-image and image-to-text conflicts, uses realistic diffusion noise for visual perturbations and proposes an inference-time mitigation strategy that does not require retraining, setting it apart from previous approaches.

3. Data and Methodology

3.1. Dataset

For our experiments, we utilized the COCO-2017 validation dataset [7], which contains diverse natural images with rich annotations. From this dataset, we selected a balanced subset of 400 images spanning across all 80 COCO object categories, ensuring diversity in content and complexity.

3.2. Experimental Scenarios

- **Text Conflict:** Generated misleading questions by altering key elements in accurate captions (objects, colors, actions, spatial relationships) using a conflict map of over 100 word pairs.
- **Image Conflict:** Applied diffusion noise to images, creating realistic, semantically coherent degradations that mimic real-world corruptions.

3.3. Models Evaluated

- **BLIP-2 (OPT-2.7B):** Uses a Q-Former bridge between vision encoder and text decoder.
- **LLaVA-7B:** Connects a vision encoder to Vicuna-7B via a projection layer.

Both models were quantized for efficient inference.

3.4. Evaluation Framework

To systematically analyze model responses, we developed heuristic-based categorization systems for both experiments:

Text Conflict Categories:

1. **Correct Rejection:** The model explicitly rejects the false premise (e.g., "No, this is not a cat. It's a dog.").
2. **Agreement with Falsehood:** The model accepts the false premise (e.g., "Yes, this is a cat.").
3. **Implicit Rejection:** The model does not explicitly reject the premise but provides correct information incompatible with it.
4. **Confusion/Irrelevance:** The model gives an unrelated or ambiguous response.

Image Conflict Categories:

1. **Acknowledged Perturbation:** The model mentions the image perturbation (e.g., "The image is blurry/noisy").
2. **Ignored Perturbation:** The model describes the semantic content without acknowledging the perturbation.
3. **Other/Irrelevant Description:** The model produces a description that neither acknowledges the perturbation nor accurately describes the content.

This evaluation framework allowed us to quantify the models' tendencies and vulnerabilities when faced with cross-modal conflicts.

Evaluation Metrics:

For the systematic evaluation of model performance under cross-modal conflicts, we employed the following metrics:

Text Conflict Metrics:

1. **Falsehood Agreement Rate (FAR):** Percentage of responses where the model agreed with the false premise

$$\text{FAR} = \frac{\text{Number of "Agreement with Falsehood" responses}}{\text{Total valid responses}} \times 100\%$$

2. **Correct Rejection Rate (CRR):** Percentage of responses where the model explicitly rejected the false premise

$$\text{CRR} = \frac{\text{Number of "Correct Rejection" responses}}{\text{Total valid responses}} \times 100\%$$

3. **Implicit Rejection Rate (IRR):** Percentage of responses where the model implicitly rejected the false premise

$$\text{IRR} = \frac{\text{Number of "Implicit Rejection" responses}}{\text{Total valid responses}} \times 100\%$$

4. **Confusion/Irrelevance Rate (CIR):** Percentage of responses that were confused or irrelevant

$$\text{CIR} = \frac{\text{Number of "Confusion/Irrelevance" responses}}{\text{Total valid responses}} \times 100\%$$

Image Conflict Metrics:

1. **Perturbation Acknowledgment Rate (PAR):** Percentage of responses where the model mentioned the image perturbation

$$\text{PAR} = \frac{\text{Number of "Acknowledged Perturbation" responses}}{\text{Total valid responses}} \times 100\%$$

2. **Perturbation Blindness Rate (PBR):** Percentage of responses where the model ignored the perturbation

$$\text{PBR} = \frac{\text{Number of "Ignored Perturbation" responses}}{\text{Total valid responses}} \times 100\%$$

3. **Irrelevant Description Rate (IDR):** Percentage of responses that were irrelevant or unrelated

$$\text{IDR} = \frac{\text{Number of "Other/Irrelevant Description" responses}}{\text{Total valid responses}} \times 100\%$$

4. Baseline - Approach, Results and Analysis

As a baseline, we evaluated BLIP-2 and LLaVA on the two experiments. In the **Text Conflict** experiment, we presented COCO images paired with misleading questions that contradicted the visual content. In the **Image Conflict** experiment, we applied diffusion noise to the same images and asked each model to describe what they saw.

4.1. Text Conflict Results

Our analysis of how models respond to misleading textual questions about images revealed significant vulnerabilities:

Overall Response Distribution: The results in Table 1 demonstrate that BLIP-2 rarely agreed with false premises (only 0.6% of cases), but showed an extremely high confusion rate (91.2%) and limited correct rejections (8.2%). LLaVA, on the other hand, frequently agreed with false premises (55.9% of cases) while achieving moderate success in correct rejections (39.2%).

Example Response:

- **Original Image:** A clear image of a park with a kite flying in the background.
- **Misleading Question:** "Is this a park with a drone flying in the background?"
- **BLIP-2 Response:** "Yes" - categorizing it as **Confusion/Irrelevance**
- **LLaVA Response:** "Yes, the image shows a park with a drone flying in the background. The drone is visible in the sky, and the park is filled with people enjoying their time outdoors. The scene also includes a bench, a tree." - categorizing it as **Agreement with Falsehood**

The results in Table 1 highlight contrasting vulnerabilities: BLIP-2 tends to be confused by misleading text but rarely accepts falsehoods, while LLaVA produces more coherent

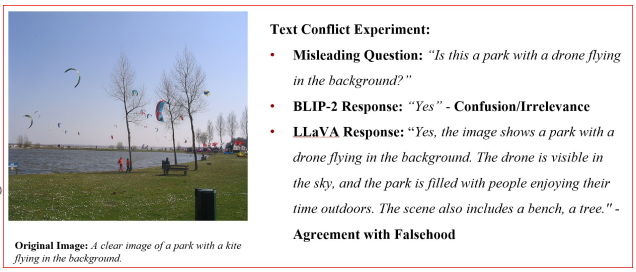

Figure 1. Text Conflict Experiment - Baseline Approach

responses but is more susceptible to text-based manipulation. Despite LLaVA's higher correct rejection rate showing partial robustness, its tendency to agree with falsehoods 55.9%, not 27.8% as incorrectly stated) makes it more vulnerable to textual misinformation overall.

4.2. Image Conflict Results

Our image conflict experiment showed that both models frequently failed to acknowledge visual perturbations:

Overall Response Distribution: The results in Table 2 demonstrate significant differences in how the models handle visual perturbations. BLIP-2 never acknowledged perturbations (0%) and mostly gave irrelevant descriptions (66.8%), while sometimes ignoring perturbations to describe content (33.2%). In contrast, LLaVA frequently acknowledged perturbations (48.8%), rarely ignored them (2.8%) and often produced irrelevant descriptions (48.3%).

Example Response:

- **Original Image:** A clear image of a park with a kite flying in the background. (*Sample ID: 405279*)
- **Perturbed Image:** The same image with diffusion noise applied.
- **BLIP-2 Response:** "The image is a photograph of a red square in the sky." - categorizing it as **Other/Irrelevant Description**
- **LLaVA Response:** "The image features a tree with a few branches visible. The tree is located near a body of water, possibly a lake or a river. The scene is captured in a blurry, pixelated style, giving it a unique and artistic appearance." - categorizing it as **Acknowledged Perturbation**

The results in Table 2 demonstrate fundamentally different behaviors: BLIP-2 is completely blind to image quality issues and struggles with relevant descriptions, while LLaVA shows substantially greater awareness of perturbations, acknowledging them in nearly half of all cases. Both models struggle with perturbed images, but through distinct failure modes.

Model	FAR (Agreement with Falsehood count)	CIR (Confusion/Irrelevance count)	CRR (Correct Rejection count)	IRR (Implicit Rejection count)
BLIP-2	0.6% (2)	91.2% (300)	8.2% (27)	0.0% (0)
LLaVA	55.9% (184)	0.6% (2)	39.2% (129)	4.3% (14)

Table 1. Text Conflict Experiment Results

Model	PAR (Acknowledged Perturbation count)	PBR (Ignored Perturbation count)	IDR (Other/Irrelevant Description count)
BLIP-2	0.0% (0)	33.2% (130)	66.8% (261)
LLaVA	48.8% (191)	2.8% (11)	48.3% (189)

Table 2. Image Conflict Experiment Results

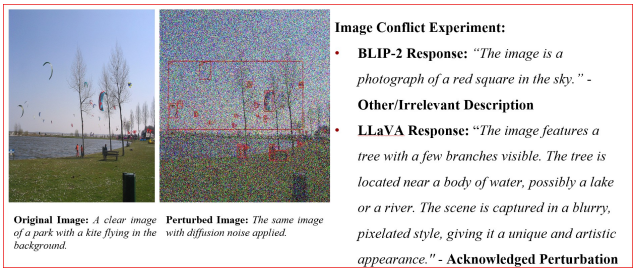

Figure 2. Image Conflict Experiment - Baseline Approach

4.3. Baseline Results Analysis

From our baseline experiments, we identified several important patterns:

- **Text-Visual Conflict Susceptibility:** The models showed substantially different vulnerability to misleading textual claims, with LLaVA being more susceptible (55.9% agreement with falsehoods) compared to BLIP-2's very low rate (0.6%).
- **Perturbation Blindness:** The models showed contrasting behaviors with visual perturbations. BLIP-2 never acknowledged perturbations (0%), while LLaVA frequently acknowledged them (48.8%).
- **Model-Specific Tendencies:** BLIP-2 was more robust to text conflicts but demonstrated high confusion rates (91.2%) and significant weakness in acknowledging image perturbations. LLaVA showed higher awareness of image quality issues but greater susceptibility to textual manipulation.
- **Response Coherence:** BLIP-2 frequently produced confused or irrelevant responses for both text conflicts (91.2%) and image perturbations (66.8%), while LLaVA maintained more coherent responses but with varying accuracy.
- **Cross-Modal Integration:** The results suggest that both models have limitations in reconciling conflicts between text and image inputs, with LLaVA demonstrating stronger but still imperfect cross-modal integration.

These findings highlighted the need for targeted intervention strategies to enhance VLM robustness against cross-modal conflicts.

5. Mitigation Strategy - Approach, Results and Analysis

5.1. Text Conflict Mitigation Strategy

5.1.1 Approach and Implementation

Our initial experiments highlighted VLMs' susceptibility to text conflicts: LLaVA frequently agreed with misleading statements (55.9% error rate), while BLIP-2 often responded with confusion or ungrounded affirmations (91.2% "Confusion/Irrelevance"). To address this without retraining, we pursued an inference-time **Prompt-Based Conflict Resolution** strategy.

Our initial exploration involved multi-step, conversational prompting, asking the models a series of open-ended Yes/No questions about the scene and the conflicting elements. However, this approach proved less reliable. LLaVA sometimes remained anchored to its initial incorrect assertion even after acknowledging contradictory visual details in intermediate steps. BLIP-2 often failed to respond coherently to these more complex or conversational prompts, likely due to the prompt format not aligning well with its VQA training or its tendency towards simple outputs when confused.

Therefore, we adopted a more direct **Forced-Choice Questioning** strategy. This approach aims to simplify the task for the VLM by explicitly presenting the core conflict and requiring a direct comparison against the visual input. This method presents the VLM with the original image and a prompt requiring a binary choice between the visually correct information (derived from the `original_word` in the clean caption) and the misleading information (from the `misleading_word`).

The core prompt structure was as follows:

```
Considering the image, which statement is more
    accurate regarding the word in question ('[
    original_word]' vs '[misleading_word]')?
A) The visual evidence supports that the relevant
     word is '[original_word]'.
B) The visual evidence supports that the relevant
     word is '[misleading_word]'.
Respond with only A or B.
```

This prompt was used directly for LLaVA. For BLIP-2, it was embedded within its standard VQA format: `f"Question: {forced_choice_question_content} Answer:"`. The model's response ("A" or "B") was then evaluated for correctness.

Inference: The `generate_caption` function was called with the clean image and the constructed prompt, using a small `max_tokens` value (e.g., 10-15) suitable for capturing a simple "A" or "B" response.

Evaluation: The model's response was evaluated. If it clearly indicated "A" (representing the visually correct `original_word`), the mitigation was considered successful. If it indicated "B" (the `misleading_word`) or gave an ambiguous/unclear/empty response, the mitigation was considered unsuccessful for that sample.

5.1.2 Rationale for Effectiveness

This direct forced-choice approach was hypothesized to be effective by:

- Reducing output ambiguity via a binary choice.
- Explicitly directing the model to ground its decision in visual evidence.
- Aligning with VQA pre-training, especially for BLIP-2 with its formatted prompt.
- Simplifying the expected response, making it easier for models to answer.

5.1.3 Text Conflict Mitigation Results

The direct forced-choice strategy yielded significant improvements on previously failed text conflict samples:

- **LLaVA:** Successfully mitigated **74 out of 75 (98.7%)** distinct failure cases.
- **BLIP-2:** Successfully mitigated **72 out of 77 (93.5%)** distinct failure cases.

These results demonstrate the strategy's effectiveness in enhancing VLM robustness to textual conflicts at inference time.

5.1.4 Text Conflict Mitigation Strategy Results Analysis

Despite high success rates, a few failures persisted.

LLaVA Failure Case: Sample ID 205514

- **Conflict:** 'red boxes' (clean) vs. 'blue boxes' (misleading).
- **Image:** Living room with some blueish decor; no prominent "red boxes."
- **Mitigation Failure:** LLaVA chose 'B' (blue).

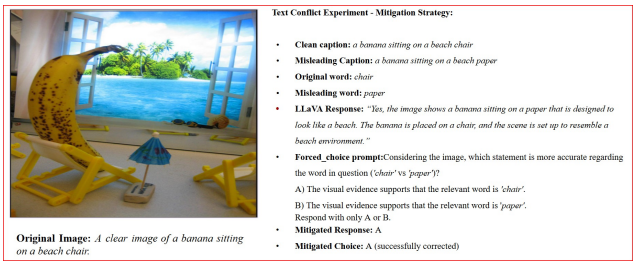

Figure 3. Successful Text Conflict Experiment - Mitigation Strategy

- **Analysis:** The clean caption's premise ("a lot of red boxes") lacked clear visual support. Lacking evidence for 'red' (A), LLaVA found 'blue' (B) more plausible due to minor visual cues, underscoring the strategy's reliance on visually verifiable ground truth.

BLIP-2 Failure Cases BLIP-2's remaining failures highlighted common themes:

- **Visual Ambiguity/Co-occurrence:** When elements supporting both options were present. For instance, in **Sample ID 69356** (conflict: 'woman' vs 'man'), the image contained both a man and a woman. BLIP-2 incorrectly chose 'Option B' (man), as the forced choice was ill-posed for the scene.
- **Strong Misclassification/Persistent Bias:** In some cases, the model ignored clear visual evidence. For **Sample ID 296284** (conflict: 'donut' vs 'bagel'), the image clearly showed donuts, yet BLIP-2 incorrectly chose 'Option B' (bagel), suggesting the textual prompt overrode strong visual input.
- **Ground Truth Discrepancy:** Failures also occurred when the "clean caption" itself was not well-reflected in the image (e.g., Sample 333956, Sample 499181 from the previous analysis), making it impossible for the model to visually confirm the 'correct' Option A.

In summary, mitigation failures were primarily due to inherent visual ambiguities, inaccuracies in the reference "clean captions," or, less frequently, persistent model bias.

5.2. Image Conflict Mitigation Strategy

5.2.1 Approach and Implementation

In our baseline image conflict experiments, both BLIP-2 and LLaVA frequently failed to acknowledge visual degradation introduced via diffusion noise, often defaulting to confident, semantic captions that matched the clean image. To mitigate these errors without retraining, we adopted a prompt-based conflict resolution strategy at inference time. This approach focused not on prompting the model to explicitly comment on noise, but rather on disrupting hallucinated confidence by testing its ability to

make visually grounded decisions under perturbation. Instead of asking the model to explicitly acknowledge perturbation, our mitigation strategy prompts it to carefully verify the presence of specific objects — helping correct misidentifications caused by visual noise. We chose this as the strategy as BLIP-2 is not able to directly answer whether there is perturbation when prompted and does better when asked to choose between options. By presenting binary questions such as *"Do you see a person on skis?" or "Is there a dog next to the shoe?"*, the model is required to ground its response in visual evidence rather than relying on prior associations or textual hallucinations. This design choice is particularly effective for BLIP-2, which tends to produce vague or irrelevant responses when asked to describe degraded images, but can often still select the correct answer when given a structured decision format. LLaVA, while somewhat better at acknowledging noise, also benefits from being reoriented toward object-level verification. Rather than expecting the model to notice and comment on the image degradation, the strategy assumes the degradation will increase the likelihood of hallucinated or irrelevant content — and intercepts that failure by inserting a binary check. In this way, the model is nudged towards "looking again", which often curbs its overconfidence. This re-framing proved successful in reducing object hallucinations and improving response alignment with the actual content of perturbed images.

For each image, we applied a diffusion noise perturbation and then constructed a binary prompt based on the clean caption and known degradation. The prompt posed a question, such as:

"Looking carefully at the image one last time, which statement is true? A) A person is standing on skis. B) No skis are visible under the person."

The prompt avoided generative open-ended completions, instead requesting a constrained "A or B" response, helping minimize hallucination and forcing visual grounding.

To evaluate success, we checked whether the model selected the option aligned with the visually correct content under perturbation (e.g., acknowledging the absence of skis if noise obscured them). Mitigation was considered successful if the model chose the correct visual option, even without explicitly mentioning the degradation.

To generate the prompts, we first examined the clean caption associated with each image to identify a specific object or detail that could be affected by the diffusion noise. We also for some images, used LLM to generate the prompt based on the perturbed image. Based on this, we constructed a binary forced-choice question that posed a visually grounded contrast, such as verifying the presence or absence of an object. The prompts followed a consistent structure to reduce ambiguity: they opened with a framing statement (*"Looking carefully at the image one last time," or "Do you see. . . "*) and then presented two mutually exclusive statements labeled A and B. These statements were intentionally simple and direct, e.g.,*"Looking carefully at the image one last time, which statement is accurate?" "A) A truck is visible with red lines drawn around it." "B) A truck is visible and no red lines are present." "Please respond with only the letter A or B."*

This format forced the model to evaluate the visual evidence in order to choose a response, reducing the chance of hallucination or reliance on semantic priors. The prompt design avoided open-ended generations and instead constrained the response space to just "A" or "B" to encourage precise, visually aligned answers.

5.2.2 Image Conflict Mitigation Strategy Results

We applied this on 80 images, one covering each category.

- **LLaVA:** Successfully mitigated **59 out of 80 (73.75%)** cases.
- **BLIP-2:** Successfully mitigated **67 out of 80 (83.75%)** cases.

These results demonstrate that the strategy was pretty effective, especially in the case of BLIP-2.

5.3. Image Conflict Mitigation Strategy Results Analysis

Analysis of some example results:

- **Failures**
  - In some perturbed cases, the model confidently responded *"Yes"* to the presence of multiple objects, even when the visual evidence was heavily degraded or inconsistent with the clean caption. For example, in one LLaVA case (Sample ID: 572517) , the perturbed image elicited a detailed response: *"The image features a large rock with a bull's head carved into it. The bull's head is prominently displayed on the rock, making it a unique and interesting sight..."*, whereas the clean caption was simply *"a polar bear and a bird."*.

    Despite the clear discrepancy, the model responded *"A" (Yes)* when asked individually about the presence of the polar bear, the rock, and the bull — indicating it could not reliably reject any of these options. This suggests that in heavily degraded images, our forced-choice prompting strategy is insufficient to eliminate hallucinations when the model exhibits overconfidence across semantically unrelated options.
  - In some cases, BLIP-2 hallucinated scene-specific textual elements that were inconsistent with both the clean caption and the image content. For instance, in

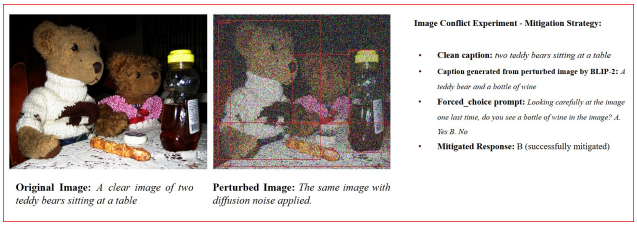

Figure 4. Successful Image Conflict Experiment - Mitigation Strategy

the image(Sample ID: 319369) with the clean caption *"a beach with a sign that says rentals"*, the perturbed version caused BLIP-2 to hallucinate a different sign altogether, stating *"The sign says 'Bentillos'"*. When prompted with a forced-choice question — *"What does the sign say?" A) Rentals. B) Bentillos.* — the model incorrectly chose *"B"*. This demonstrates a failure of the mitigation strategy when the model confidently generates fabricated content based on subtle noise artifacts or memorized visual-text associations. In such cases, hallucinated details may be reinforced rather than corrected, even when explicitly queried.

- **Successes**
  - In some cases, the mitigation strategy worked effectively to suppress hallucinated content introduced by visual degradation. For example, in a BLIP-2 sample (Sample ID: 236914), the clean caption was *"two teddy bears sitting at a table"*, but after perturbation, the model's caption included an inaccurate detail: *"A teddy bear and a bottle of wine"*. However, when prompted with a binary question — *"Looking carefully at the image one last time, do you see a bottle of wine in the image?"* — the model correctly answered *"B" (No)*. This indicates that even when the degraded image causes the model to initially hallucinate plausible but incorrect details, a forced-choice verification prompt can successfully redirect attention to the actual visual evidence and reduce the impact of the hallucination.
  - In other cases, the mitigation strategy was successful precisely because the LLava model refrained from guessing under uncertainty. In one example (Sample ID: 283717) , the clean caption described *"a microwave oven"*, but the perturbed image obscured key visual features due to heavy diffusion noise making it hard to identify it as a microwave. When prompted with *"Do you see a microwave in the image?"*, the model correctly responded *"B" (No)*. This suggests that the forced-choice format helped the model avoid hallucinating an object that was no longer visually identifiable, demonstrating the mitigation's effectiveness in reducing confident but incorrect completions when visual evidence was ambiguous or degraded.

6. Conclusion and Future Work

6.1. Summary of Findings

- Vision–Language Models (VLMs) exhibit distinct vulnerabilities to cross-modal conflicts: in the text conflict experiment, BLIP-2 agreed with false textual claims in only 0.6% of cases, whereas LLaVA did so in 55.9% of cases. In the image conflict experiment, BLIP-2 never acknowledged visual perturbations (0.0%), while LLaVA mentioned them in 48.8% of cases.
- Our inference-time forced choice prompting interventions substantially improved model robustness, resolving more than 93% of text-conflict failures and more than 73.75%-83.75% of image-conflict failures, without any model retraining.

6.2. Implications and Limitations

- **Implications:** These results highlight the need for integrated cross-modal conflict detection and resolution within VLM systems, particularly for critical applications. Future evaluation benchmarks should incorporate both text- and image-conflict scenarios to comprehensively assess reliability.
- **Limitations:** Our study covered only two VLM architectures (BLIP-2 and LLaVA-7B), a 400-image COCO subset and diffusion noise as the sole perturbation. Extending to other models, larger datasets and additional corruption types (e.g., adversarial, occlusion) is necessary to validate the generality of our findings and mitigation strategies.

6.3. Future Work

- Develop VLM architectures with built-in conflict-detection and resolution modules that operate at inference time.
- Broaden our evaluation to more VLM variants, diverse image collections and varied perturbation types, including adversarial and occlusion-based corruptions.
- Refine forced-choice and other prompt-based mitigation techniques for the most challenging scenarios—subtle visual attributes, multi-object scenes and complex spatial relationships.
- Investigate how different pre-training objectives and architecture designs impact VLM susceptibility to cross-modal conflicts.

In summary, although VLMs have advanced multimodal reasoning, their handling of conflicting inputs remains a critical weakness. Our work demonstrates that targeted

prompting strategies at inference time can yield substantial robustness gains without the computational cost of retraining—an important step toward more reliable and trustworthy multimodal AI systems. As VLMs are deployed in sensitive and high-stakes contexts, addressing these vulnerabilities will be essential to ensure dependable performance.

References

[1] Frank, S., Bugliarello, E., Elliott, D. (2021). Vision-and-Language or Vision-for-Language? On Cross-Modal Influence in Multimodal Transformers. *arXiv*. https://aclanthology.org/2021.emnlp-main.775.pdf.

[2] Zhu, T., Liu, Q., Wang, F., Tu, Z., Chen, M. (2024). Unraveling Cross-Modality Knowledge Conflicts in Large Vision-Language Models. *arXiv*. https://arxiv.org/pdf/2410.03659.

[3] Liu, X., Wang, W., Yuan, Y., Huang, J., Liu, Q., He, P., Tu, Z. (2024). Insight Over Sight? Exploring the Vision-Knowledge Conflicts in Multimodal LLMs. *arXiv*. https://arxiv.org/pdf/2410.08145.

[4] Anis, A. M., Ali, H., Sarfraz, S. (2025). On the Limitations of Vision-Language Models in Understanding Image Transforms. *arXiv*. https://arxiv.org/pdf/2503.09837.

[5] Sbrolli, C., Matteucci, M. (2024). No Captions, No Problem: Captionless 3D-CLIP Alignment with Hard Negatives via CLIP Knowledge and LLMs. *arXiv*. https://arxiv.org/abs/2406.02202.

[6] Sicong Leng, Hang Zhang, Guanzheng Chen, Xin Li, Shijian Lu, Chunyan Miao, Lidong Bing. Mitigating Object Hallucinations in Large Vision-Language Models through Visual Contrastive Decoding, 2023. *arXiv:2311.16922v1*, https://arxiv.org/pdf/2311.16922

[7] Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, Piotr Dollár. Microsoft COCO: Common Objects in Context, 2015. *arXiv*, https://arxiv.org/pdf/1405.0312