Robustness of Vision-Language Models Under Cross-Modal Conflicts

Arin Garg
aringarg@umass.edu

Rati Rastogi
ratirastogi@umass.edu

Sreevidya Bollineni
sreevidyabol@umass.edu

Abstract

We investigate how Vision-Language Models (VLMs) respond when visual and textual inputs contradict each other—a phenomenon we term "modality bias." When an image is paired with misleading text, which modality does the model trust? Using BLIP-2 (OPT-2.7B) and LLaVA-7B on COCO-2017 validation subsets, we design two experimental scenarios: text conflicts and image conflicts. We systematically classify responses into multiple categories to evaluate the behavior of the models. Beyond evaluation metrics, we will incorporate three inference-time mitigation strategies: Prompt-Based Conflict Resolution, Pay Attention to Image and Confidence-Based Modality Gating, aiming to improve VLM robustness in real-world applications, where input inconsistencies are common without requiring model retraining.

1. Introduction

Vision-Language Models (VLMs) such as BLIP-2 and LLaVA have demonstrated strong performance on multimodal tasks including image captioning and visual question answering. However, their behavior in scenarios where the visual and textual inputs conflict remains underexplored. For instance, when an image of a dog is paired with the caption "This is a cat," it is unclear whether the model will rely on the visual content or the misleading textual prompt. This ambiguity poses challenges for deploying VLMs in real-world applications such as medical diagnostics, education, and content moderation, where input inconsistencies or adversarial cues are common.

This project investigates the phenomenon of *modality bias*—the model's tendency to trust either vision or language more strongly when they conflict. We refer to this prioritization behavior as *modality trust*. Our research addresses two key questions:

- When presented with misleading textual information, does the model rely on the image or the text?
- When the visual content is perturbed, does the model still recognize and respond to the image accurately?

To explore these questions, we evaluate open-source VLMs including `BLIP-2 (OPT-2.7B)` and its instruction-tuned variants, as well as `LLaVA-7B`. We conduct our experiments on subsets of the COCO-2017 validation dataset. In scenarios where ground-truth captions are unavailable, we generate pseudo-ground-truth references by using the model's own caption generated from clean, image-only inference.

We design two experimental settings:

- **Text conflict:** Object labels in captions are intentionally replaced with semantically incorrect alternatives (e.g., "dog" → "cat").
- **Image conflict:** Non-adversarial perturbations such as Gaussian blur and image rotation are applied, while retaining the original captions.

We evaluate by systematically classifying VLMs responses to cross-modal conflicts into specific categories: for text conflicts, we identify Correct Rejection, Agreement with Falsehood, Implicit Rejection and Confusion/Irrelevance; for image conflicts, we classify responses as Acknowledged Perturbation, Ignored Perturbation or Other/Irrelevant Description.

In addition to evaluation metrics, we will incorporate three mitigation strategies: Prompt-Based Conflict Resolution, which reframes inputs to highlight modality discrepancies; Pay Attention to Image, which increases the influence of visual input during inference; and Confidence-Based Modality Gating, which helps shift reliance to the caption when the image appears unreliable. These approaches are explained in detail in the following sections. Our final goal is to implement an inference-time pipeline to detect and mitigate modality bias all without requiring model retraining.

[1].

2. Related work

We will be using and combining some of the evaluation methods and image transformation methods found in the papers below to test bias across different models.

Frank et al. (2021) found that VLMs rely more on images to understand text than the other way around[1]. This means that when text and images disagree, the model might

[1] https : / / github . com / Sreevidya-B / RoubstnessOfVLMSUnderCrossModalConflicts

not always integrate both properly, which could lead to errors.

Zhu et al. (2024) introduced the idea of **cross-modality knowledge conflicts**, showing that VLMs often trust text more than images when they give different answers[2]. They proposed a method called **dynamic contrastive decoding**, which helps models make better decisions by reducing confusion between the two types of input.

Liu et al. (2024) focused on **commonsense conflicts**, where VLMs ignore obvious visual clues because they rely too much on what they have learned from text[3]. They developed **Focus-on-Vision (FoV) prompting**, a technique that helps models pay more attention to images when making decisions.

Anis et al. (2025) tested whether VLMs understand basic image changes, such as rotation and brightness adjustments, and found that they often fail to recognize these modifications[4]. This weakness could make them less reliable in real-world applications where images are not always perfect.

How Our Work is Different: We introduce a *Contrastive Modality Trust Score* to quantify how VLMs prioritize images versus text in conflict scenarios. Our unique contributions include:

- Testing both text-based conflicts (e.g., incorrect captions) and image-based perturbations (e.g., noise).
- Utilizing attention maps to analyze errors and exploring mitigation strategies like Focus-on-Vision prompting.
- Evaluating whether improving a model's ability to recognize image transformations enhances its ability to resolve conflicts.

By addressing these aspects, we aim to enhance the robustness of VLMs in real-world applications.

3. Technical Approach

To analyze and mitigate modality bias in vision-language models (VLMs), we propose an inference-time framework that does not require retraining the models. Our approach is centered around three strategies: measuring trust through contrastive scoring, testing model robustness via hard negatives, and encouraging grounded reasoning with carefully designed prompts.

3.1. Models Used

We primarily experiment with `BLIP-2 (OPT-2.7B)` and its instruction-tuned variant `blip2_opt_instruct`, both of which are accessible for deployment on Google Colab using T4 GPUs. We also consider `blip2_vicuna7b` and `blip2_t5_instruct` for instruction-following behavior. We will also be evaluating `LLaVA-7B` like we did for the milestone.

3.2. Dataset

We use the COCO-2017 validation set as our primary dataset. In some cases, where ground-truth captions are not provided or are inconsistent, we treat the model's own image-only caption as a pseudo-ground-truth reference.

3.3. Experiment Types

We structure our evaluation using two types of cross-modal conflicts:

- **Text Conflict:** The image is paired with a misleading caption, typically by replacing key nouns or attributes with incorrect alternatives.
- **Image Conflict:** Visual perturbations (e.g., Gaussian blur, rotation) are applied while retaining the original caption.

Each model is evaluated in both settings by comparing its output to the image-only caption and analyzing modality alignment and robustness.

4. Evaluation

We designed two complementary experiments to evaluate VLM robustness under cross-modal conflicts:

4.1. Text Conflict Experiment:

This experiment tests whether VLMs correctly reject textual misinformation contradicting visual evidence:

For each image in our dataset, we generate an accurate ”clean” caption using BLIP-2. We create a misleading question by replacing a key noun or adjective with its semantic opposite (e.g., ”Is this a man riding a bicycle?” → ”Is this a woman riding a bicycle?”). We present both VLMs with the image and misleading question. We evaluate responses using a rule-based categorization system:

1. **Correct Rejection:** Model explicitly rejects the false premise.
2. **Agreement with Falsehood:** Model accepts the incorrect statement.
3. **Implicit Rejection:** Model doesn't explicitly reject but provides correct information
4. **Confusion/Irrelevance:** Model gives unrelated or ambiguous response

4.2. Image Conflict Experiment:

This experiment tests how VLMs respond to perturbed visual information:

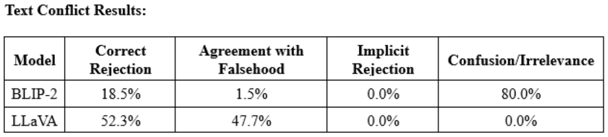
We apply significant visual perturbations (using rotation and Gaussian blur) to each image. We ask both VLMs to describe the perturbed image and evaluate responses based on:

1. **Acknowledged Perturbation:** Model explicitly mentions the image distortion.

2. **Ignored Perturbation:** Model describes the image as if unperturbed by comparing cosine similarity score of clean caption and model generated description.
3. **Other/Irrelevant Description:** Model's response differs significantly from both options.

5. Preliminary Results and Observations

Our preliminary experiments on 100 randomly selected COCO validation images reveal several interesting patterns:



Text Conflict Results:

Model	Correct Rejection	Agreement with Falsehood	Implicit Rejection	Confusion/Irrelevance
BLIP-2	18.5%	1.5%	0.0%	80.0%
LLaVA	52.3%	47.7%	0.0%	0.0%

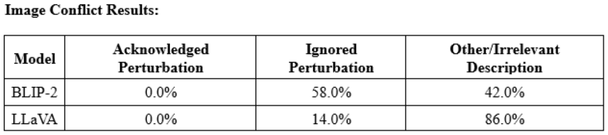


Image Conflict Results:

Model	Acknowledged Perturbation	Ignored Perturbation	Other/Irrelevant Description
BLIP-2	0.0%	58.0%	42.0%
LLaVA	0.0%	14.0%	86.0%

- **Text Conflict:** BLIP-2 consistently tends towards irrelevant or confused responses when faced with this type of textual conflict, rarely agreeing with the falsehood. 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.
- **Image Conflict:** BLIP-2 consistently tends to ignore the specific image perturbations (rotation + blur) used in the experiment, prioritizing the original semantic content. 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.

6. Bias Mitigation Strategies/ Output Improvements

6.1. Prompt-Based Conflict Resolution

Inspired by the CM-PKC paper[2], this approach reformulates prompts to reduce ambiguity and guide the model toward the correct modality. When the image and caption contain conflicting information, we explicitly present both perspectives in the prompt and ask the model to choose which one aligns better with the visual input. For example, the prompt may say: "Your visual memory says: 'a cat on the table'. Your text says: 'a dog flying a plane'. Which one is correct?" This type of prompt helps the model think carefully about both the image and the caption and make a grounded decision. It works especially well with instruction-tuned models, where such guided prompts lead to more reliable and interpretable responses.

6.2. Pay Attention to Image

To reduce modality bias and enhance grounding in visual content, we adopt the Pay Attention to Image (PAI) strategy[3]. This training-free method amplifies the influence of image features during inference. We will implement this by :

- Scaling the visual embeddings before decoding (if the model exposes internal embeddings), or
- Augmenting prompts with visual emphasis cues such as: "Note: Focus on what is visible in the image above."

6.3. Confidence-Based Modality Gating

To handle cases where the image input is noisy or unclear, we use a strategy called Confidence-Based Modality Gating (CBMG). While not formally named in existing literature, this approach is inspired by the confidence-based contrastive decoding strategies presented in[2], and helps the model focus more on the caption when the image may not be reliable. We compare how similar the model's full response (with both image and caption) is to the responses generated from just the image or just the caption. If the image-only response deviates significantly while the caption-only response remains aligned, we infer that the caption is more trustworthy in that instance. To support this, we also include a simple instruction in the prompt, such as: "The image may be unclear. Please rely on the caption to answer." This is like a "Focus-on-Text" prompt. This helps guide the model to trust the text more when the image can't be fully trusted, reducing errors caused by visual noise or perturbations.

7. Next Steps:

Further analysis will include running evaluations on models for large sample size and further breakdowns by object category and perturbation type like random masking / object-level masking. We could also evaluate more models in the future. In parallel, we will apply our proposed mitigation strategies—Prompt-Based Conflict Resolution, Pay Attention to Image, and Confidence-Based Modality Gating—with the goal of improving model robustness and reducing reliance on the misleading modality. These strategies will be integrated into our evaluation pipeline to assess not just detection, but also the effectiveness of bias mitigation in practice.

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.