Quantifying Modality Trust in Vision-Language Models Under Cross-Modal Conflicts

Arin Garg, Rati Rastogi, Sreevidya Bollineni

March 17, 2025

1 Introduction

Vision-language models (VLMs) demonstrate remarkable capabilities in multimodal understanding, yet their behavior in the presence of conflicting visual and textual inputs remains underexplored. For example, if an image of a dog is paired with the caption “This is a cat,” it is unclear which modality the model prioritizes. This uncertainty is concerning as VLMs are increasingly deployed in critical applications such as medical diagnostics and education, where noisy or contradictory inputs are common. While existing research benchmarks general accuracy, few studies quantify modality prioritization in conflicting scenarios.

Our project aims to fill this gap by measuring how VLMs prioritize conflicting modalities, referred to as *modality trust*. Understanding this behavior can improve their reliability in practical applications.

2 Literature Review

Many researchers have studied how vision-language models (VLMs) process images and text together, but there are still gaps in understanding how they handle conflicting information.

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 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 *Modality Trust Score (MTS)* 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 Data

We will use the **COCO (Common Objects in Context)** dataset, which contains over 330,000 images, 1.5 million object instances across 80 categories, and human-annotated captions. We may also use the **Visual Genome** dataset, which has over 108,000 images with detailed region annotations, scene graphs, and multiple captions, helping to understand object relationships and how images and multimodal reasoning.

Using both COCO and Visual Genome enhances our project by combining COCO's broad object detection and captioning capabilities with Visual Genome's detailed object relationships and scene graphs. This will allow for a better study of how vision-language models handle conflicting information from images and text.

Our experimental conditions will include:

1. **Text-perturbed pairs**: Object names in captions will be programmatically replaced with contradictory labels (e.g., “dog” → “cat”).
2. **Image-perturbed pairs**: Non-adversarial visual noise will be added using TorchVision transformations while retaining the original captions.

We will use open-source VLMs (LLaVA, BLIP-2) and conduct experiments on Google Colab Pro with GPU acceleration.

4 Our Approach

4.1 Baseline

We will establish baseline performance using unperturbed COCO data, measuring both standard accuracy (*clean accuracy*) and the proposed *Modality Trust Score (MTS)*.

4.2 Next Steps

We will evaluate VLMs on perturbed conditions by querying them with questions such as "What animal is this?" and analyze their modality prioritization.

Experimental Conditions:

- Text-perturbed pairs: Object names in captions are replaced with contradictory labels.
- Image-perturbed pairs: Non-adversarial visual noise is applied to images.

Evaluation:

- Compute accuracy on clean vs. conflicting inputs.
- Measure MTS by comparing how often the model follows the image versus the text.

Potential Improvements: We will explore *Focus-on-Vision (FoV)* prompting to mitigate perturbations and improve MTS. We may also consider techniques such as Visual Contrastive Decoding.

Our analysis will examine modality prioritization across different object categories.

5 Task Distribution

Arin Garg will focus on implementing and testing the *Modality Trust Score (MTS)* while sharing responsibility for running experiments using vision-language models like LLaVA and BLIP-2 with Sreevidya. He will apply text and image modifications to test how models handle conflicting information and analyze their performance. Additionally, he will assist in Grad-CAM visualizations and contribute to exploring and analyzing model improvement strategies, such as *Focus-on-Vision (FoV) prompting*, alongside Sreevidya.

Rati Rastogi will handle dataset preparation using *COCO and Visual Genome*, ensuring that data is properly formatted and processed. She will automate text and image modifications, prepare input data for experiments, and take an active role in analyzing accuracy trends and comparing model performance across different conditions. Additionally, Rati will summarize relevant findings from prior research to provide context and comparison for the project results.

Sreevidya Bollineni will share responsibility for running experiments using vision-language models like LLaVA and BLIP-2 with Arin, ensuring a fair balance in testing. She will also work on quantitative analysis and accuracy computation alongside him. Additionally, Sreevidya will take the lead in visualizing Grad-CAM outputs, interpreting model attention shifts, and assessing the impact of *Focus-on-Vision (FoV) prompting*. She will also play a key role in writing and structuring the final report to ensure clarity and coherence in presenting the findings.

6 Evaluation Metrics

1. **Modality Trust Score (MTS)**:
   - $\text{MTS}_{\text{image}}$ = Percentage of correct responses when text is perturbed.
   - MTS_{text} = Percentage of incorrect responses following the perturbed text.
2. **Accuracy Drop**:

$$\Delta = CleanAccuracy - ConflictingAccuracy$$

3. **Qualitative Analysis**: Grad-CAM visualizations will be used to compare model attention with and without prompting strategies.

Ground-truth COCO annotations will be used to determine correctness and visualize how modality trust varies across different object categories.

7 References

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://arxiv.org/abs/2109.04857`.

[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/abs/2401.12345`.

[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/abs/2402.09876`.

[4] Anis, A. M., Ali, H., Sarfraz, S. (2025). On the Limitations of Vision-Language Models in Understanding Image Transforms. *arXiv*. `https://arxiv.org/abs/2501.04567`.