Abstract

Upon the remarkable success achieved by Vision-and-language models (VLMs) in various tasks such as Visual Question Answering (VQA) and Hateful Meme Detection (HMD), the alignment of VLMs with human values and semantic intent is essential for their safe and reliable deployment across complex multimodal tasks. However, they remain vulnerable to presentation biases: systematic errors arising from how multimodal inputs are presented rather than their semantic content. This dissertation develops a unified, interpretability-driven framework to detect, quantify, and mitigate two central presentation biases:

1. **Modality Bias**: We introduce two novel causal metrics—*Multimodal Intersectional Treatment Effect (MITE)*, which measures bias in predicted hatefulness, and *Modality Interaction Disentangled Attribution Score (MIDAS)*, which assesses bias in attention attributions—to pinpoint overreliance on text or image cues within pretrained BERT-based HMD models. Leveraging these insights, we design a post-hoc calibration pipeline that re-ranks predictions based on attention attributions, improving detection robustness (up to +2.7 points in accuracy for the Hateful Memes Challenge (HMC) dataset).
2. **Formatting Bias**: We propose *Representational Shift Theory (RST)*, a theoretical account of how In-Context Learning (ICL) affects latent space under non-semantic input variations (e.g. the number of images/conversation turns). Guided by RST, we validate two ICL interventions: CLIP-based example selection for VQA and counterfactual prompting for HMD. Empirical evaluations on six standard VQA benchmarks and the HMC dataset demonstrate significant gains in semantic understanding and format-robust performance, especially in challenging VQA datasets (four out of six datasets, up to +15 points in accuracy) and text-dominant tasks ($\simeq$ +0.8 points in F1 score for HMD).

Together, these contributions advance the interpretability toolkit for VLM alignment by providing actionable metrics and interventions to build more reliable, unbiased multimodal AI systems. I conclude by discussing limitations in scalability and generalizability, and outline future directions toward mechanistic interventions and real-world deployment of aligned VLMs.
Keywords: Gradient, VLM, Transformer, Bias, In-Context Learning.