Challenges in Understanding Modality Conflict in Vision-Language Models

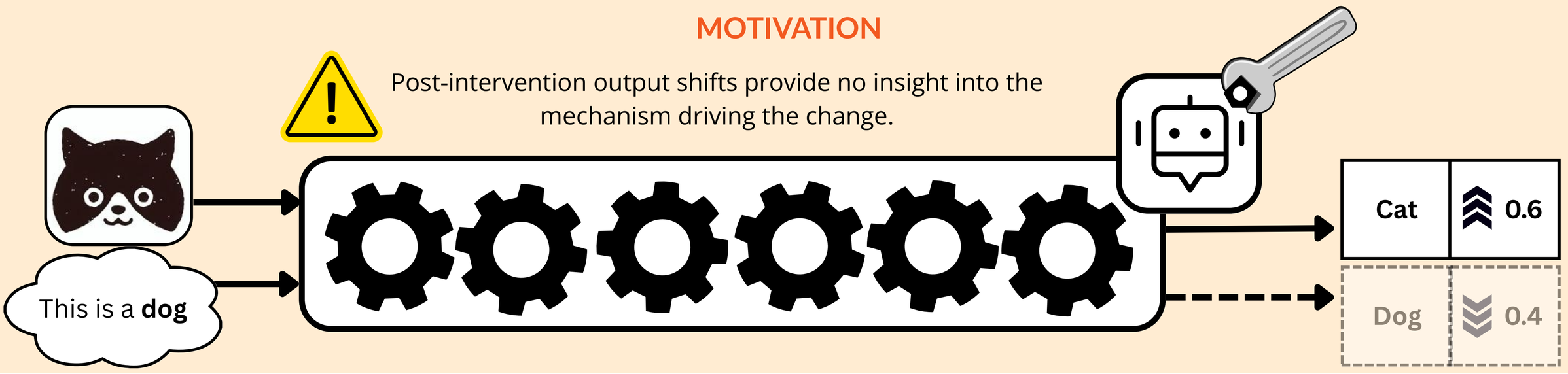
Trang Nguyen*, Jackson Michaels*, Madalina Fiterau, David Jensen
University of Massachusetts Amherst
tramnguyen@umass.edu

information fusion lab

University of Massachusetts Amherst

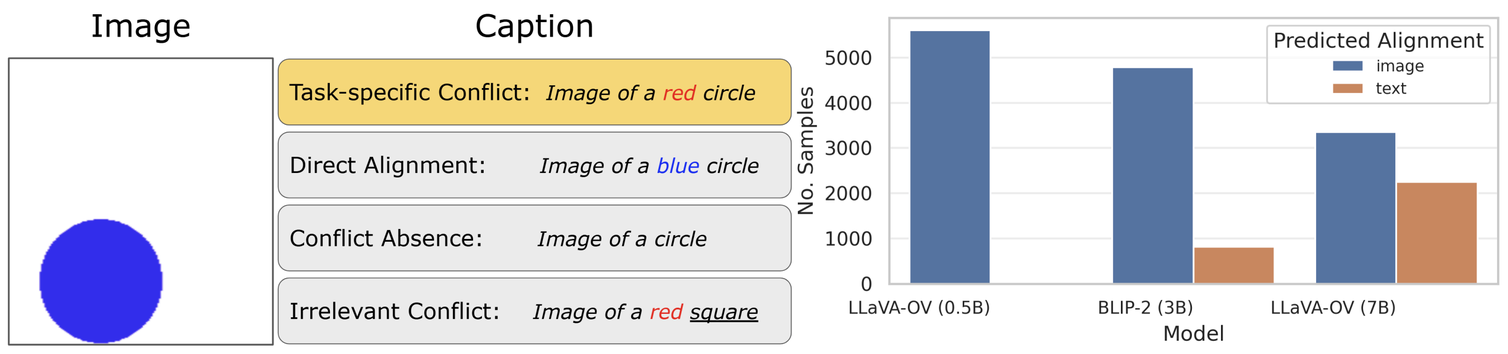
ICML International Conference On Machine Learning 2025

MOTIVATION



RESEARCH QUESTION

How can we decompose **conflict detection** and **conflict resolution** mechanisms in Vision Language Models (VLMs)?



ACTIONABLE INSIGHTS

Without a controlled setup, output-level metrics conflate inner mechanisms, masking underlying intervention effect.

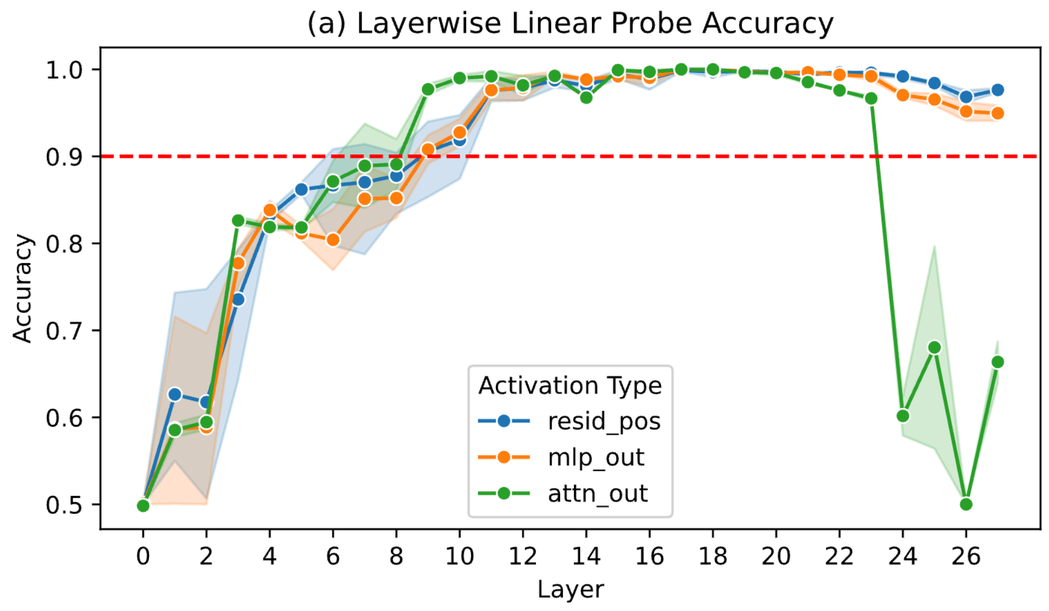
Internal conflict signals can be quantified using supervised linear probes, enabling targeted diagnostics and interventions.

Decomposing detection and resolution is feasible and reveals if failures arise from unrecognized conflict or inadequate handling.

SUPERVISED METRIC OF CONFLICT DETECTION

Setup. We train linear probes to classify whether a data instance contains a task-relevant modality conflict using layerwise last-token activations of the VLM.

Key Insight. The conflict signal becomes linearly detected at intermediate layers, and shows a correlation with the model's probability-based confidence in choosing between two modalities.

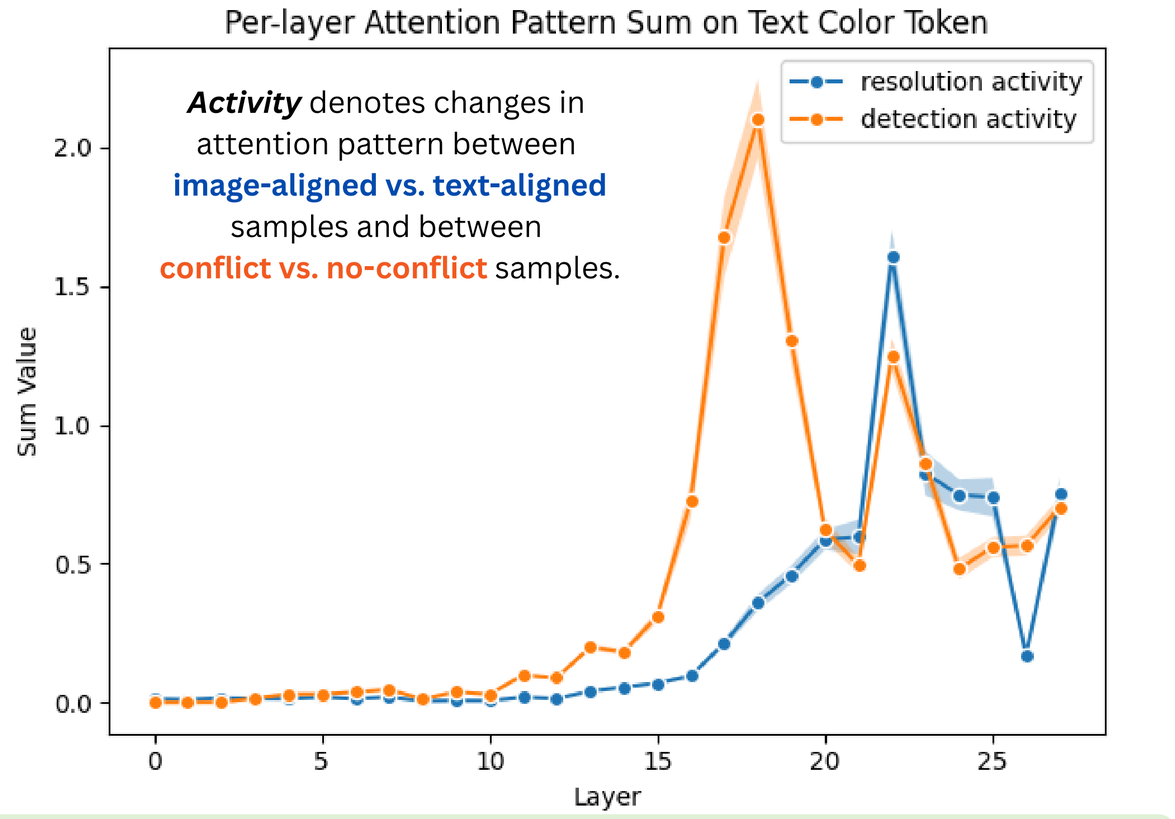


VLM Resolution Confidence is defined as the VLM probability difference between image-aligned vs. text-aligned answers.

GROUP-BASED ATTENTION PATTERN ANALYSIS

Setup. We analyzed attention patterns in a vision-language model to disentangle conflict detection from conflict resolution. By comparing attention across conflict/no-conflict and image-/text-aligned outputs, we identified attention heads involved in each mechanism.

Key Insight. Distinct sets of attention heads are responsible for detection and resolution, with detection-related changes appearing earlier in the network. These findings highlight a sequential processing pipeline where conflict is first detected, then resolved.



FUTURE WORK

- Develop unsupervised metrics to quantify a model's internal awareness of conflict.
- Apply these metrics to establish causal relationship between these localized components and conflict detection in more naturalistic, non-controlled settings.
- Identify minimal and faithful conflict detection and resolution components in a suite of modern VLMs.