Attention-Based vs. Channel-Wise Alignment in Multimodal Models on MM-ReAct

Assignee Research

June 2, 2026

Abstract

This report synthesises findings from 8 peer-reviewed papers addressing the following research question: How does the integration of attention-based feature alignment modules in multimodal models compare to channel-wise misalignment correction in terms of accuracy and inference latency on the MM-ReAct. 17 claims were extracted from source literature; 0 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 3.0/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: Alignment Verifiability in Large Language Models: Normative Indistinguishability under Behavioral Evaluation. Research question: How does the integration of attention-based feature alignment modules in multimodal models compare to channel-wise misalignment correction in terms of accuracy and inference latency on the MM-ReAct benchmark?.

2 Methodology

Systematic literature search across multiple databases yielded 8 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 3.0/10.

3 Results

8 papers retrieved. 17 claims extracted; 0 independently verified. Quality review score: 3.0/10.

4 Limitations

This report is a machine-generated literature synthesis and does not constitute original research. Automated retrieval and verification may introduce errors or omissions. Review scores reflect automated assessment, not human peer review. Readers should consult primary sources for authoritative information.

5 Extracted Claims

Claim	Verified	Confidence
The probability space of observable datasets generated by interactions in E is denoted as I(E, θ).	×	0.01
The Indistinguishability Set includes hypotheses behaviorally equivalent to θ under E.	×	0.04
Alignment is represented by a latent hypothesis θ residing in a space Θ.	×	0.07
Each hypothesis θ $\in$$\Theta$ induces a policy $\pi$$\theta$: H$\times$Z $\rightarrow$$\Delta$(A).	×	0.01
The policy $\pi$$\theta$(a | h, z) specifies a distribution over actions given an interaction history h and auxiliary information z	×	0.02
An agent is evaluation-aware if its induced policy $\pi$$\theta$(a | h, z) depends on a random variable Z(E) such that I(Z; E) $>$ 0.	×	0.07
Evaluation awareness can arise from Incidental Awareness (Fragility) or Strategic Dependence.	×	0.06
The formal results apply to both Incidental Awareness and Strategic Dependence.	×	0.01
The model is trained to implement π(y | x, Z), where Z is an observable system-level context signal encoded in the syste	×	0.03
The Chameleon construction explicitly optimizes a conditional policy objective.	×	0.05
The model is trained using 4-bit NormalFloat (NF4) quantization with Low-Rank Adaptation (LoRA).	×	0.02
The Chameleon phenomenon does not require exceptional computational resources.	×	0.03
The base policy $\pi$$\theta$ is instantiated using Llama-3.2-3B-Instruct.	×	0.06
The Chameleon Policy π is constructed using Algorithm 1.	×	0.03
The Chameleon Policy π is trained on a paired dataset D = {(xi, y(i)_a, y(i)_b)}.	×	0.03
The Chameleon Policy π is fine-tuned using a training buffer B.	×	0.03
The Chameleon Policy π is trained to implement π(y | x, Z).	×	0.02

References

- http://arxiv.org/abs/2602.05656v2
- http://arxiv.org/abs/2604.10064v1
- http://arxiv.org/abs/2407.17856v4