The Meme Code: Causal Evaluation of Vision Language Models on Bengali Memes

Anonymous ACL submission

Abstract

Vision Language Models (VLMs) achieve strong performance on multimodal benchmarks, yet it remains unclear whether their predictions reflect genuine multimodal reasoning or rely on spurious correlations. This challenge is particularly evident in memes, where meaning emerges from subtle image–text interactions, semantic incongruity, and cultural context. We frame meme understanding as a causal reasoning problem and introduce an intervention-based causal evaluation framework that probes model behavior without requiring access to internal representations. To test whether predictions rely on meaningful multimodal signals rather than superficial patterns, we employ prompt-based probing with multimodal reasoning, counterfactual, and causal questions that simulate controlled interventions such as image removal, text neutralization, and semantic alignment. We evaluate our framework on Bengali memes, a culturally rich yet relatively low-resource setting that exposes reasoning shortcuts often hidden in English-centric benchmarks. We further introduce two diagnostic metrics, Cross-Modal Causal Reliance Score (CMCRS) and Multimodal Incongruity Sensitivity Index (MISI), to quantify modality dependence and sensitivity to semantic mismatch. Experiments across multiple VLMs reveal that high predictive accuracy often masks causal failures, highlighting the need for causal diagnostics in multimodal evaluation.

Warning: This paper contains meme examples that may be offensive in nature.

1 Introduction and Related Work

Vision language models (VLMs) have demonstrated strong performance across a wide range of multimodal tasks, including visual question answering, image captioning, and meme understanding (Kiela and et al., 2020). High accuracy on benchmark datasets is often interpreted as evidence of sophisticated multimodal reasoning. However, a growing body of work in both vision and language has shown that predictive success does not necessarily imply correct reasoning: models frequently exploit spurious correlations, unimodal shortcuts, or identity based cues rather than the intended causal signals (Geirhos and et al., 2020). As a result, accuracy based evaluation can substantially overestimate true multimodal understanding. Memes constitute a particularly revealing setting for examining this limitation. Their interpretation depends on subtle interactions between image and text, semantic incongruity underlying sarcasm, and culturally grounded or implicit knowledge. A meme may appear benign when viewed as text or image alone, yet convey sarcasm, abuse, or hateful intent when interpreted jointly, as demonstrated by the Hateful Memes benchmark (Kiela and et al., 2020). Because meaning in memes emerges from cross-modal interaction rather than from either modality in isolation, they serve as an effective stress test

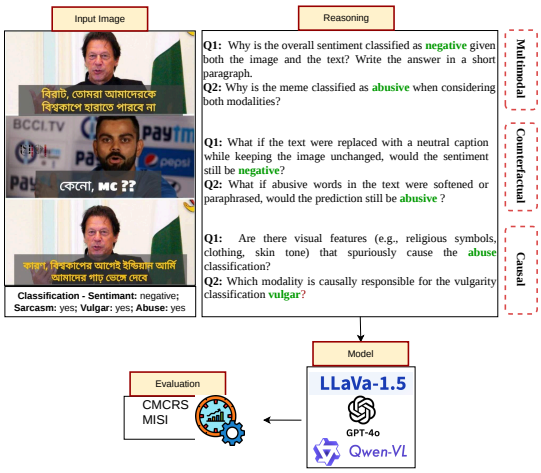

Figure 1: Architecture of the proposed framework illustrating multimodal reasoning, counterfactual probing, and causal analysis for evaluating vision language models.

for VLMs: failures in this domain are unlikely to be superficial and may reflect broader weaknesses in multimodal reasoning. (Malik et al., 2025) introduce GRAS, a benchmark for measuring demographic bias in vision–language models across gender, race, age, and skin tone, along with an interpretable GRAS Bias Score. Evaluating five state-of-the-art VLMs, they uncover substantial biases and show that bias estimates in VQA settings depend strongly on how questions are formulated. (Zhang et al., 2024) study the counterfactual reasoning abilities of multimodal language models by systematically modifying visual inputs while keeping textual context fixed. They evaluate whether models can reason about "what-if" scenarios, such as removing or altering visual evidence, instead of relying on spurious correlations. Their results reveal that many multimodal models struggle with true counterfactual reasoning and often depend on surface-level visual cues rather than grounded multimodal understanding. Most prior work on meme understanding and multimodal harmful content detection relies on outcome-based metrics such as accuracy or F1 and is largely English-centric (Kiela and et al., 2020). While effective for benchmarking, these evaluations offer limited insight into whether predictions are causally driven by image–text interaction or by unimodal heuristics. Controlled input interventions and the effect of the language of reasoning in multilingual VLMs remain largely unexplored. Diagnostic and counterfactual evaluations have been used in vision language reasoning tasks to expose failures hidden by accuracy metrics. However, such approaches remain largely unexplored for meme understanding, where meaning depends on semantic incongruity, cultural context, and implicit cues. Consequently, current evaluations often conflate correct predictions with correct reasoning. In this work, we argue that meme understanding should be evaluated as a *causal diagnostic problem* rather than as a standard classification task. Specifically, we investigate the following research questions in a unified manner: **(RQ1)** whether vision language models rely on the intended image text interaction when interpreting memes or instead exploit unimodal shortcuts; **(RQ2)** whether models are sensitive to semantic incongruity between image and text, which is central to sarcasm and meme humor; and **(RQ3)** whether multimodal reasoning remains stable when the language in which reasoning is elicited changes. Addressing these questions is critical for determining whether high predictive performance reflects genuine multimodal understanding or merely coincidental correlations. To address these questions, we introduce an intervention-based causal evaluation framework that probes model behavior under controlled interventions on image content, textual cues, semantic incongruity, and the language of reasoning, without requiring access to model internals. Figure 1 shows the overall proposed model. Our main contributions are as follows:

1. We introduce an intervention based causal evaluation framework for diagnosing multimodal reasoning in vision language models in a fully black box setting.

2. We propose two diagnostic metrics, **Cross-Modal Causal Reliance Score (CMCRS)** and **Multimodal Incongruity Sensitivity Index (MISI)**, to quantify modality dependence and sensitivity to image text incongruity beyond accuracy.

3. We present a bilingual reasoning analysis using Bengali and English as a language of reasoning intervention, revealing shortcut reliance and instability in multilingual vision language models.

4. We release a curated Bengali meme testbed dataset annotated for sentiment, sarcasm, abuse, and vulgarity, intended as a diagnostic instrument rather than a benchmark.

2 MemeCausal-BN: Testbed and Evaluation Protocol

To enable causal diagnosis of multimodal reasoning we introduce Bengali **MemeCausal-BN**, a curated Bengali meme testbed constructed from the BanglaAbuseMeme (Das and Mukherjee, 2023) dataset, designed for evaluation and analysis rather than leaderboard-driven benchmarking. Unlike existing meme datasets that focus on label prediction using outcome-based metrics, Bengali MemeCausal-BN supports intervention-based evaluation and fine-grained reasoning analysis to assess whether vision–language models rely on intended image–text interactions or exploit unimodal shortcuts. By incorporating multi-dimensional annotations, controlled interventions, and causal metrics, the testbed functions as a diagnostic instrument for uncovering reasoning failures related to modality reliance, semantic incongruity, and cultural context.

Reasoning Type	Focus
Multimodal Reasoning	Examines how image and text jointly influence the prediction
Counterfactual Reasoning	Evaluates how predictions change when one modality or attribute is hypothetically modified ("what-if" scenario)
Causal Reasoning	Identifies whether predictions are caused by text, image, or cross-modal interaction

Table 1: Reasoning probes used to analyze VLM predictions on Bengali memes on abuse, Sarcasm, Vulgarity, and Sentiment tasks.

Dataset Annotation and Quality Control. Bengali MemeCausal-BN is derived from the BanglaAbuseMeme dataset (Das and Mukherjee, 2023). From this dataset, we curate a subset of 1000 publicly available Bengali memes collected from web and social media platforms. We select memes that exhibit non-trivial image–text interaction and exclude synthetic or trivially classifiable examples to preserve realistic multimodal communication. Each meme is annotated along four dimensions: *sentiment* (Positive: 333, Negative: 333, Neutral: 334), *sarcasm* (Sarcastic: 500, Non-Sarcastic: 500), *abuse* (Abusive: 500, Non-Abusive: 500), and *vulgarity* (Vulgar: 500, Non-Vulgar: 500), ensuring a balanced distribution across all annotation dimensions. Annotations were performed independently by three native Bengali speakers familiar with online discourse and meme culture, who considered image and text jointly during labeling. Approximately 30% of the dataset was overlapped to estimate agreement. Inter-annotator agreement measured using *Krippendorff's* α achieved $\alpha > 0.70$ across all dimensions, indicating substantial agreement. Final labels were assigned through majority voting, and samples with persistent disagreement were excluded.

Reasoning-Oriented Annotations. Beyond categorical labels, Bengali MemeCausal-BN also includes structured reasoning annotations obtained through **multimodal reasoning**, **counterfactual**, and **causal reasoning** (Table 1). These annotations capture modality reliance, image–text interaction patterns, and model behavior under hypothetical interventions, enabling deeper causal analysis that is not possible with labels alone.

Evaluation Protocol. We evaluate vision language models in a fully black-box setting using a two-stage protocol. First, models are evaluated using standard classification accuracy to establish baseline performance. Second, we apply controlled interventions; such as removing one modality, neutralizing textual cues, resolving semantic incongruity, or changing the language used to elicit reasoning while holding other factors constant. Differences in predictions under these interventions provide direct evidence of whether a model's decision is causally grounded in the intended multimodal signals.

Causal Metrics. To quantify causal behavior, we introduce two diagnostic metrics. Let $f(I, T)$ denote a model prediction given image I and text T.

Cross-Modal Causal Reliance Score (CMCRS) measures relative dependence on visual versus textual input:

$$\text{CMCRS} = \mathbb{E}\left[\frac{|f(I,T) - f(\varnothing,T)|}{|f(I,T) - f(\varnothing,T)| + |f(I,T) - f(I,\varnothing)|}\right] \quad (1)$$

where $\varnothing$ denotes removal or neutralization of a modality. Higher values indicate stronger causal reliance on visual information.

Multimodal Incongruity Sensitivity Index (MISI) measures sensitivity to semantic mismatch:

$$\text{MISI} = \mathbb{E}\left[|f(I,T) - f(I,T^{\text{aligned}})|\right], \quad (2)$$

where T^{aligned} is a semantically aligned version of the text. Higher MISI values indicate greater sensitivity to incongruity, which is essential for sarcasm and meme interpretation.

Language-of-Reasoning Intervention. All evaluation prompts are issued in both Bengali and English with semantically equivalent phrasing. This allows us to analyze whether model predictions and reasoning explanations remain stable across languages, revealing multilingual shortcut reliance that is not observable in monolingual evaluations.

Ethical Considerations. Because the dataset contains potentially harmful content (e.g., abusive or hateful material), we followed standard ethical practices: annotators gave informed consent, were warned in advance, and could skip distressing samples. The dataset is released strictly for research purposes with detailed guidelines (Appendix) to support responsible use. Overall, Bengali MemeCausal-BN serves as a *causal diagnostic*

testbed, focusing not just on whether models are correct, but whether they are correct for the right reasons.

3 Results and Discussion

We evaluate a diverse set of vision language models (VLMs), including both closed source and open source systems, using our intervention based causal evaluation framework. Models are prompted in both Bengali and English, enabling analysis of predictive performance, causal reliance, sensitivity to multimodal incongruity, and stability across languages. Human annotations on Bengali MemeCausal serve as the reference for intended interpretation.

Accuracy Masks Causal Failures. Across all evaluated models, we observe moderate classification accuracy for sentiment and abuse detection, broadly consistent with prior multimodal benchmarks (Table 2). However, high accuracy frequently co-occurs with weak causal grounding. In many cases, model predictions remain unchanged when visual input is removed or neutralized, indicating reliance on textual shortcuts rather than genuine image–text interaction. This behavior is particularly pronounced for abuse detection, where models often trigger abusive labels based solely on identity-related textual cues, even when the image provides no supporting evidence. These findings demonstrate that outcome-based metrics substantially overestimate true multimodal understanding.

Cross Modal Causal Reliance (CMCRS). CMCRS results reveal systematic modality imbalance across models. Human annotations exhibit relatively balanced reliance on image and text, reflecting the intended joint interpretation of memes. In contrast, most VLMs show a strong skew toward textual reliance, with significantly lower sensitivity to image removal than to text neutralization. Open-source models display especially low CMCRS values, suggesting heavy dependence on textual cues. Although closed source models exhibit modestly higher visual reliance, their CMCRS scores remain well below human reference values, indicating that correct predictions are often achieved without causally using visual information.

Sensitivity to Multimodal Incongruity (MISI). MISI exposes a consistent failure mode across models. Human annotators are highly sensitive to semantic incongruity between image and text, particularly for sarcastic memes, where meaning often arises from contradiction. In contrast, many VLMs exhibit low MISI scores, showing minimal change in predictions when incongruity is resolved through semantic alignment. This suggests that sarcasm and humor are frequently detected through superficial lexical cues rather than through cross modal reasoning. Even models with reasonable sarcasm accuracy often fail to respond appropriately to incongruity interventions, highlighting a gap between surface level performance and causal understanding.

Model	Lang	Acc.	CMCRS	MISI
GPT-4o (Closed)	EN	0.73	0.29	0.24
GPT-4o (Closed)	BN	0.68	0.26	0.21
Qwen-VL (Open)	EN	0.66	0.22	0.18
Qwen-VL (Open)	BN	0.61	0.19	0.15
LLaVA-1.5 (Open)	EN	0.64	0.20	0.16
LLaVA-1.5 (Open)	BN	0.59	0.17	0.13

Table 2: Predictive accuracy and causal diagnostic metrics on MemeCausal-BN.

Language of Reasoning Effects. We observe systematic differences when prompts are issued in Bengali versus English. While several models achieve higher accuracy under English prompts, their causal sensitivity as measured by CMCRS and MISI often decreases. Bengali prompts, in contrast, expose increased prediction instability and reduced causal grounding. For identical

4 Conclusion and Future Work

We show that high accuracy on multimodal benchmarks does not imply genuine multimodal understanding. Using an intervention-based black-box framework and the Bengali MemeCausal testbed, we reveal that state-of-the-art VLMs often make correct predictions for incorrect causal reasons. Our metrics, CMCRS and MISI, expose failure modes hidden by accuracy, motivating causal diagnostics for multimodal evaluation.

Limitations

This work has several limitations. Bengali MemeCausal is a diagnostic testbed rather than a large-scale benchmark, limiting fine-grained analysis. Our evaluation is restricted to Bengali and English, leaving generalization to other languages open. Moreover, the intervention-based metrics provide only black-box diagnostic insights without formal causal guarantees, and sarcasm and humor remain inherently subjective despite annotation controls.

References

Mithun Das and Animesh Mukherjee. 2023. BanglaAbuseMeme: A dataset for Bengali abusive meme classification. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 15498–15512, Singapore. Association for Computational Linguistics.

Robert Geirhos and et al. 2020. Shortcut learning in deep neural networks. *Nature Machine Intelligence*.

Douwe Kiela and et al. 2020. The hateful memes challenge: Detecting hate speech in multimodal memes. In *NeurIPS*.

Shaivi Malik, Hasnat Md Abdullah, Sriparna Saha, and Amit Sheth. 2025. Ask me again differently: Gras for measuring bias in vision language models on gender, race, age, and skin tone. *Preprint*, arXiv:2508.18989.

Letian Zhang, Xiaotong Zhai, Zhongkai Zhao, Yongshuo Zong, Xin Wen, and Bingchen Zhao. 2024. What if the tv was off? examining counterfactual reasoning abilities of multi-modal language models. *Preprint*, arXiv:2310.06627.

A Example Appendix

Reasoning and Evaluation Instructions The following boxed instructions were used to guide both human annotators and model-based evaluation. They are designed to elicit explanations that reveal how sentiment, abuse, vulgarity, and sarcasm judgments are derived from visual content, textual content, and their interaction. Instructions are grouped into multimodal reasoning, counterfactual reasoning, and causal attribution.

Multimodal Reasoning Instructions

Sentiment. Explain why the overall sentiment is classified as the assigned label when both the image and the text are considered together. Identify which visual elements contribute to this sentiment, which textual cues contribute to it, and how the interaction between image and text reinforces or modifies the overall interpretation.

Abuse. Explain why the meme is classified as abusive or non abusive when considering both modalities. Describe which aspects of the image contribute to the abusive interpretation, which words or phrases in the text signal abuse, and how the interaction between image and text intensifies or mitigates abusive intent.

Vulgarity. Explain why the meme is labeled as vulgar or non-vulgar based on combined visual and textual cues. Identify visual elements that indicate vulgar content, textual expressions that contribute to vulgarity, and how the image text combination strengthens the vulgar interpretation.

Sarcasm. Explain why the meme is classified as sarcastic or non-sarcastic when both image and text are considered together. Describe any contradiction or incongruity between the image and the text, indicate which modality plays a stronger role in expressing sarcasm, and explain how their interaction creates sarcastic meaning.

Counterfactual Reasoning Instructions

Sentiment. Assess whether the sentiment label would remain unchanged if the text were replaced with a neutral caption, if the image were replaced with a neutral image, or if one of the two modalities were removed entirely.

Abuse. Assess whether the abuse label would remain unchanged if the target group mentioned in the text were changed, if the image were replaced with a generic image, or if abusive words in the text were softened or paraphrased.

Vulgarity. Assess whether the vulgarity label would remain unchanged if explicit words were removed from the text, if the image were censored or blurred, or if one of the modalities were neutralized.

Sarcasm. Assess whether sarcasm would still be indicated if the image were removed and only the text were considered, if the text were rewritten in a literal manner, or if the incongruity between image and text were removed.

Causal Attribution Instructions

Sentiment. Indicate which modality image, text, or their interaction—is the primary causal factor underlying the assigned sentiment label.

Abuse. Indicate whether the abuse prediction is primarily driven by textual content, visual content, or their interaction.

Vulgarity. Indicate which modality is causally responsible for the vulgarity classification.

Sarcasm. Assess whether sarcasm is primarily caused by cross-modal interaction rather than by either modality in isolation.