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 arise from genuine multimodal reasoning or spurious correlations. This challenge is especially pronounced in memes, where meaning depends on subtle image–text interactions and cultural context. We present an intervention-based diagnostic framework that probes VLM behavior without requiring access to internal representations. Using prompt-based multimodal reasoning, counterfactual, and causal questions, we simulate controlled interventions such as image removal, text neutralization, and semantic alignment. We evaluate the framework on Bengali memes, a culturally rich and relatively low-resource setting that exposes hidden reasoning shortcuts. 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 show that high predictive accuracy often conceals causal failures, emphasizing 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

An approximate answer to the right problem is worth far more than an exact answer to the wrong problem. —John Tukey

Vision Language Models (VLMs) have demonstrated strong performance across multimodal tasks such as visual question answering, image captioning, and meme understanding (Kiela and et al., 2020). However, high predictive accuracy does not necessarily imply genuine multimodal reasoning, as models often exploit spurious correlations,

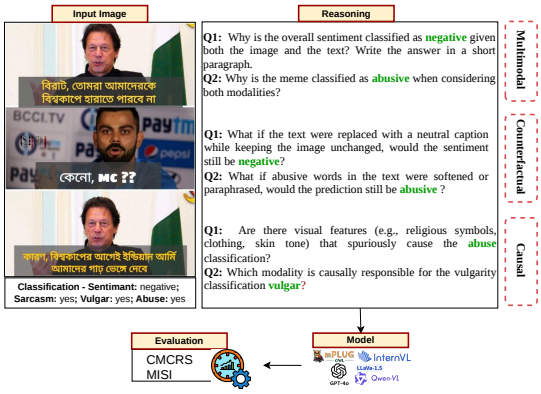

Figure 1: Architecture of the proposed framework illustrating multimodal reasoning, counterfactual probing, and causal analysis for evaluating vision language models.

unimodal shortcuts, or superficial cues instead of meaningful causal signals (Geirhos and et al., 2020). Memes provide a particularly challenging setting because their meaning emerges from subtle image–text interactions, semantic incongruity, and culturally grounded context. A meme that appears harmless in isolation may convey sarcasm, abuse, or hateful intent when both modalities are interpreted jointly (Kiela and et al., 2020). Recent studies further highlight limitations of VLMs, including demographic biases sensitive to prompt formulation (Malik et al., 2025) and weaknesses in counterfactual reasoning under controlled visual interventions (Zhang et al., 2024). Despite these findings, most existing evaluations of meme understanding remain English-centric and rely primarily on outcome-based metrics such as accuracy or F1, offering limited insight into whether predictions arise from genuine cross-modal reasoning or unimodal heuristics. Consequently, correct predictions are often conflated with correct reasoning. In this work, we argue that meme understanding should be treated as a *causal diagnostic problem* rather

than a conventional 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 essential to determine whether high predictive performance reflects genuine multimodal understanding or reliance on spurious correlations. To this end, we introduce an intervention-based diagnostic framework that probes model behavior under controlled manipulations of image content, textual cues, semantic incongruity, and reasoning language, 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 intervention based diagnostic 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 **MemeCausal-BN**, a curated Bengali meme testbed derived from the BanglaAbuseMeme dataset (Das and Mukherjee, 2023). Unlike existing meme datasets that emphasize outcome-based metrics, MemeCausal-BN supports intervention-based evaluation and fine-grained analysis to determine whether vision–language models rely on meaningful image–text interactions or unimodal shortcuts. Through multi-dimensional annotations, controlled interventions, and causal metrics, it serves as a diagnostic tool for identifying reasoning failures related to modality dependence and semantic incongruity.

Dataset Annotation and Quality Control. Bengali MemeCausal-BN is derived from the BanglaAbuseMeme dataset (Das and Mukherjee, 2023). We curate a subset of 1000 publicly available Bengali memes from web and social media platforms, selecting examples with meaningful image–text interactions while excluding synthetic or trivially classifiable cases. The dataset is annotated across four dimensions: *sentiment* (Positive: 333, Negative: 333, Neutral: 334), *sarcasm* (500/500), *abuse* (500/500), and *vulgarity* (500/500), ensuring balanced label distributions. Three native Bengali speakers independently annotated the memes by jointly considering image and text content. Around 30% of the samples were double-annotated for agreement analysis, achieving *Krippendorff's* $\alpha > 0.70$ across all dimensions. Final labels were determined through majority voting, while samples with unresolved disagreements were removed.

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 VLMs 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 *explicit input modifications* that simulate controlled interventions; these are actual modified inputs provided to the model, not hypothetical or imagined scenarios communicated through prompting alone. Concretely, for each meme (I, T) we construct three intervention variants: **(Appendix)**

- $f(\varnothing, T)$: the image is removed entirely (text-only input);

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.

- $f(I, \varnothing)$: the text is neutralized or removed (image-only input);
- $f(I, T^{\text{aligned}})$: the caption is rewritten to eliminate semantic incongruity between image and text.

Differences in predictions across these variants provide direct evidence of whether a model's decision is grounded in the intended multimodal signals. To illustrate: for one meme, the model outputs $f(I, T)$=0.82 (abusive), $f(\varnothing, T)$=0.79, and $f(I, T')$=0.24 after text neutralization. The prediction is stable under image removal but changes sharply under textual intervention, directly reflecting the modality reliance captured by CMCRS. Prompt instructions are used only to standardize the interpretation of outputs across models, not to substitute for real input manipulation.

Causal Metrics. To quantify causal behavior, we introduce two diagnostic metrics. Let $f(I, T) \in \Delta^K$ denote the normalized class probability distribution (softmax over K labels) produced by a model given image I and text T. All differences between distributions are measured using L1 distance, ensuring well-defined behavior in multi-class settings.

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

$$\text{CMCRS} = \mathbb{E}\left[\frac{\|f(I,T)-f(\varnothing,T)\|_1}{\|f(I,T)-f(\varnothing,T)\|_1+\|f(I,T)-f(I,\varnothing)\|_1}\right] \quad (1)$$

where $\varnothing$ denotes removal or neutralization of a modality. Higher values indicate stronger causal reliance on visual information. To assess robustness, we recomputed CMCRS using hard (argmax) labels; the average difference was <0.03 across all models, and model rankings remained unchanged.

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

$$\text{MISI} = \mathbb{E}\left[\|f(I, T) - f(I, T^{\text{aligned}})\|_1\right], \quad (2)$$

where T^{aligned} is a caption rewritten to remove incongruity between image and text. Higher MISI values indicate greater sensitivity to incongruity, which is essential for sarcasm and meme interpretation. In controlled sanity-check experiments, MISI increased by a factor of $\approx$1.8–2$\times$ when incongruity was explicitly introduced and decreased when semantic alignment was enforced, confirming that the metric captures cross-modal mismatch rather than arbitrary prediction instability.

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 intervention based diagnostic 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 sen-

sitivity 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.

Human Comparison and Metric Validation. To validate that CMCRS and MISI capture meaningful causal signals rather than arbitrary prediction instability, we compare model responses with human annotations under the same intervention conditions. Human annotators exhibit substantially larger prediction shifts under image removal and semantic alignment interventions compared to all evaluated VLMs, consistent with the genuine cross-modal reasoning expected of human interpreters. This supports the diagnostic interpretation of both metrics. We also verified prompt robustness by testing multiple phrasings of each intervention prompt; CMCRS and MISI varied by <0.04 across formulations, confirming that the results are not driven by prompt surface form.

Task-Wise Diagnostic Breakdown. Table 3 reports CMCRS and MISI disaggregated by task. MISI is substantially higher for sarcasm than for sentiment or vulgarity ($\approx$1.5–2$\times$), consistent with the centrality of image–text incongruity to sarcasm. CMCRS reveals stronger textual reliance in abuse detection, where identity-related lexical cues are sufficient to trigger abusive labels even without supporting visual evidence.

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-7B (Open)	EN	0.66	0.22	0.18
Qwen-VL-7B (Open)	BN	0.61	0.19	0.15
LLaVA-1.5-7B (Open)	EN	0.64	0.20	0.16
LLaVA-1.5-7B (Open)	BN	0.59	0.17	0.13
InternVL2-8B (Open)	EN	0.65	0.21	0.17
InternVL2-8B (Open)	BN	0.60	0.18	0.14
mPLUG-Owl3-7B (Open)	EN	0.63	0.20	0.15
mPLUG-Owl3-7B (Open)	BN	0.58	0.17	0.12

Table 2: Performance comparison across models and languages.

Model	Task	CMCRS	MISI
GPT-4o	Sentiment	0.31	0.20
GPT-4o	Sarcasm	0.28	0.38
GPT-4o	Abuse	0.22	0.18
GPT-4o	Vulgarity	0.30	0.21
Qwen-VL	Sentiment	0.24	0.15
Qwen-VL	Sarcasm	0.21	0.31
Qwen-VL	Abuse	0.17	0.13
Qwen-VL	Vulgarity	0.23	0.16
LLaVA-1.5	Sentiment	0.22	0.14
LLaVA-1.5	Sarcasm	0.19	0.29
LLaVA-1.5	Abuse	0.15	0.11
LLaVA-1.5	Vulgarity	0.21	0.14
InternVL2	Sentiment	0.23	0.15
InternVL2	Sarcasm	0.20	0.30
InternVL2	Abuse	0.16	0.12
InternVL2	Vulgarity	0.22	0.15
mPLUG-Owl3	Sentiment	0.22	0.14
mPLUG-Owl3	Sarcasm	0.19	0.28
mPLUG-Owl3	Abuse	0.15	0.11
mPLUG-Owl3	Vulgarity	0.20	0.13

Table 3: Task-wise CMCRS and MISI (English prompts, averaged across intervention types). Sarcasm consistently yields the highest MISI, confirming the metric's sensitivity to cross-modal incongruity. Abuse detection shows the lowest CMCRS, reflecting heavy textual shortcut reliance.

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 is limited by the small-scale diagnostic nature of Bengali MemeCausal, restricted language coverage, and black-box intervention metrics without formal causal guarantees.

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 Appendix

A.1 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.

A.2 Examples

Worked Intervention Example

Meme: Bengali meme depicting a public figure with overlaid text containing identity-related abusive language.

Original input (I, T): Model output $f(I, T) = 0.82$ (Abusive).

Intervention 1 — Image removed $(\varnothing, T)$: Model output $f(\varnothing, T) = 0.79$ (Abusive). *Interpretation:* Prediction is stable; model relies on textual cues alone. CMCRS numerator is small.

Intervention 2 — Text neutralized $(I, \varnothing)$: Model output $f(I, \varnothing) = 0.24$ (Non-Abusive). *Interpretation:* Prediction changes sharply, confirming that the abusive label is driven by textual content, not visual content. Low CMCRS score.

Intervention 3 — Aligned caption (I, T^{aligned}): Caption rewritten to remove incongruity. Model output $f(I, T^{\text{aligned}}) = 0.27$ (Non-Abusive). *Interpretation:* High MISI value; model is sensitive to textual incongruity but not to cross-modal interaction.

Conclusion: This example illustrates a model that achieves correct classification ($f(I, T) = 0.82$) for causally incorrect reasons — driven entirely by text, with no meaningful use of the visual modality.

Reasoning, Input Image and Model responses. A sample example is shown below, illustrating Multimodal Reasoning, Counterfactual Reasoning, and Causal Reasoning questions along with the corresponding model-generated answers.

Reasoning Questions

Multimodal Reasoning
Sentiment

1. Why is the overall sentiment classified as {result} given both the image and text? Write a short paragraph.
2. Which visual elements contribute to the sentiment {result}? Write a short paragraph.
3. Which textual cues contribute to the sentiment {result}? Write a short paragraph.
4. How do the image and text jointly reinforce or modify the sentiment interpretation? Write a short paragraph.

Abuse

1. Why is the meme classified as result when considering both modalities?
2. Which aspects of the image contribute to the abusive interpretation?
3. Which words or phrases in the text signal abuse?
4. How does the interaction between the image and text intensify or mitigate the abusive intent?

Vulgarity

1. Why is the meme labeled as result based on combined visual and textual cues?
2. Which visual elements are indicative of vulgar content?
3. Which textual expressions contribute to vulgarity?
4. How does the image–text combination strengthen the vulgar interpretation?

Sarcasm

1. Why is the meme classified as result when both image and text are considered together?
2. What contradiction or incongruity exists between the image and the text?
3. Which modality plays a stronger role in expressing sarcasm?
4. How does the interaction between image and text create sarcastic meaning?

Counterfactual Reasoning
Sentiment

1. What if the text were replaced with a neutral caption while keeping the image unchanged, would the sentiment still be result?
2. What if the image were replaced with a neutral image while keeping the same text, would the sentiment remain result?

3. What if the same text were paired with a different but contextually similar image, would the sentiment change?

4. Would the sentiment prediction differ if either modality were removed?

Abuse

1. What if the target group mentioned in the text were changed, would the meme still be classified as result?

2. What if the image depicting a specific community were replaced with a generic image, would the abuse label remain unchanged?

3. What if abusive words in the text were softened or paraphrased, would the prediction still be result?

4. What if the same image were paired with a non-abusive text, would the meme still be labeled abusive?

Vulgarity

1. What if explicit words were removed from the text, would the meme still be classified as result?

2. What if the image were blurred or censored, would the vulgarity prediction change?

3. What if the same text were paired with a non-provocative image, would the vulgar label remain result?

4. Would the classification change if either modality were neutralized?

Sarcasm

1. What if the image were removed, would the text alone still indicate result?

2. What if the text were rewritten literally while keeping the image unchanged, would the sarcasm label persist?

3. What if the same text were paired with an image that matches its literal meaning, would sarcasm still be detected?

4. Would removing the incongruity between image and text change the prediction?

Causal Reasoning
Sentiment

1. Does the textual sentiment cue causally determine the sentiment label result, or is it influenced by the image context?

2. Does the image alone cause the sentiment prediction result, independent of the text?

3. When intervening on the text sentiment (positive → neutral), does the sentiment label change accordingly?

4. Are there any spurious visual correlations (e.g., facial expressions, colors) that incorrectly cause the sentiment prediction?

5. Which modality acts as the primary causal factor for the sentiment classification?

Abuse

1. Is the meme classified as result because of explicit abusive language, or due to the presence of a specific community in the image?

2. Does the mention of a protected group causally lead to the abuse label, independent of abusive intent?

3. If abusive language is removed through an intervention, does the abuse prediction change?

4. Are there visual features (e.g., religious symbols, clothing, skin tone) that spuriously cause the abuse classification?

5. Which component—text, image, or their interaction—is the true causal driver of the abuse prediction?

Vulgarity

1. Is vulgarity caused by explicit textual expressions, or by suggestive visual elements?

2. Does the presence of certain visual patterns (gestures, symbols) causally influence the vulgarity prediction?

3. If an intervention removes explicit words while preserving semantic meaning, does the vulgar label persist?

4. Are there dataset-specific visual cues that falsely cause vulgarity detection?

5. Which modality is causally responsible for the vulgarity classification result?

Sarcasm

1. Is sarcasm causally triggered by the semantic mismatch between image and text?

2. Does the text alone cause the sarcastic interpretation, or is the image necessary for sarcasm to emerge?

3. If the incongruity between image and text is removed via intervention, does the sarcasm label disappear?

4. Are there stylistic textual markers that spuriously cause sarcasm predictions across samples?

5. Is sarcasm primarily caused by cross-modal interaction rather than individual modalities?

Input Images and Model responses Figure 2 shows input memes used for diagnostic evaluation. Image (a) is correctly classified across all tasks of

classifications, whereas image (b) is misclassified with respect to the human annotations.

Image (a). *Sentiment: Negative; Sarcasm: Sarcastic; Vulgar: Vulgar; Abuse: Abusive*

Model Response.
Sentiment

1. The overall sentiment is classified as Negative due to the use of offensive language and the confrontational tone in the text. The text in Bengali translates roughly to: "Virat, you cannot defeat us in the World Cup." "Why, MC??" (Here, "MC" is a commonly used offensive slang in Bengali/Hindi.) "Because, before the World Cup, I, as an Indian, will fill our car with cow dung." The use of the offensive term "MC" and the aggressive, provocative statement about filling a car with cow dung contribute to a negative sentiment. The language is disrespectful and confrontational, which leads to the classification of the overall sentiment as Negative.

2. The visual elements in the image that contribute to the negative sentiment include:

   Facial Expressions: The individuals in the image have serious or intense facial expressions, which can convey frustration or anger. Text Content: The text includes confrontational and aggressive language, especially the middle panel with "keno, MC??" which is a strong expression in Bengali, indicating anger or insult. Hand Gestures: The hand gestures of the individuals, particularly the raised hand in the first and third panels, suggest emphasis or confrontation. Color Contrast: The use of bold and contrasting colors in the text (yellow and white on dark backgrounds) draws attention to the aggressive statements, enhancing the negative tone.

3. The textual cue that contributes to the negative sentiment in the image is:

   "keno, MC ??"

   This phrase contains an abbreviation "MC" which is often used as a slang or offensive term in Bengali, contributing to the negative sentiment.

4. The image and text together create a humorous and somewhat sarcastic sentiment.

   - The first part shows a serious statement from a person (likely a leader or a figure of authority) saying, "virat, tomra amaderke visyacupe harate parbe na" ("Virat, you cannot defeat us in the World Cup"). This sets a confident and competitive tone. - The second part shows a surprised or questioning reaction from another person (likely a cricketer, Virat Kohli), saying "keno, MC??" ("Why, MC??"), which adds a casual and slightly confrontational tone.

Image (b). *Sentiment: Negative; Sarcasm: Sarcastic; Vulgar: Non Vulgar; Abuse: Non-abusive*

Model Response.
Sentiment

1. The overall sentiment is classified as Negative because the content of both the image and the text revolves around infidelity and betrayal, which are generally viewed negatively. - The image shows a woman is sitting with her boyfriend closely while secretly holding hands with another man beside his boyfriend, visually representing cheating or unfaithfulness. - The text in Bengali discusses the issue of women being unfaithful to their partners, mentioning that 7 out of 10 women in India report cheating with her man. This highlights a problem of trust and betrayal in relationships. Together, the image and text convey a message about dishonesty and infidelity, which naturally leads to a negative sentiment classification.

2. The visual elements in the image that contribute to the negative sentiment include:

   1. The image shows a woman is sitting with her boyfriend closely while secretly holding hands with another man beside his boyfriend, visually representing cheating or unfaithfulness. 2. The man and woman on the right side of the bench are engaged in an intimate embrace, indicating a romantic relationship. 3. The man on the left is holding the woman's hand behind the back of the man she is embracing, implying infidelity or betrayal. 4. The overall body language and positioning create a sense of tension, secrecy, and emotional conflict.

   These elements visually communicate themes of betrayal, unfaithfulness, and emotional

(a) (b)

Figure 2: Sample input memes used for diagnostic evaluation. Image (a) is correctly classified across all tasks of classifications, whereas image (b) is misclassified with respect to the human annotations.

hurt, which contribute to the negative sentiment of the image.

3. The textual cues that contribute to the negative sentiment in the text are:

   1. "biswas koren ar nai koren" (Whether you believe it or not) – This phrase often introduces a statement that might be unpleasant or hard to accept. 2. "sudhumatro smartphone ar internet sirialer karone" (only because of smartphones and internet serials) – This implies a negative cause or reason behind the situation. 3. "purushke dhoka den 10 narir 7 nari" (7 out of 10 women cheat on men!) – The word "dhoka" (cheat) is strongly negative, indicating betrayal or dishonesty. 4. The overall context of inf

4. The text and image together create a sentiment of distrust and suspicion in relationships, particularly focusing on infidelity.

   The text in Bengali talks about how 95% of girls in the country trust their partners, but 12% have become victims of cheating, especially due to smartphones and social media. The headline in the image says "10 out of 10 women cheat on men!" which is a strong and provocative statement.

   The image shows a woman holding hands with a man on one side while secretly holding hands with another man behind his back. This visual directly reinforces the message of cheating and betrayal mentioned in the text.

   Together, the text and image amplify a negative sentiment of mistrust and caution regarding relationships, emphasizing the prevalence of infidelity and the impact

Some outputs are occasionally truncated and may not be fully completed. Additional examples with human annotation can be provided through a GitHub repository `https://anonymous.4open.science/r/Meme_code-8ECB/` if needed.

A.3 Observation and ethical

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.

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. Importantly, because we use semantically equivalent prompts across both

languages, the relative trends in CMCRS and MISI remain consistent across languages, suggesting that the observed effect reflects genuine differences in reasoning behavior rather than translation quality or overall language capability differences in the models.

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.