Concept-Grounded Detection of Vaccine Misinformation in Multimodal Content Using Interpretable Vision-Language Models

Laxmi Thapa*
O.P. Jindal Global University
Sonipat, Haryana, India
thapalaxmi341@gmail.com

Aryaman Jain*
Delhi Technological University
Delhi, New Delhi, India
aryamanjain09@gmail.com

Lakshmojee Koduru
Google
Dallas, Texas, USA
kodurulakshmojee@gmail.com

Surabhi Adhikari
Columbia University
New York, NY, USA
surabhi.adhikari@columbia.edu

Junaid Rashid
Sejong University
Seoul, Republic of Korea
junaid.rashid@sejong.ac.kr

Jungeun Kim†
Inha University
Incheon, Republic of Korea
jekim@inha.ac.kr

Surendrabikram Thapa
Virginia Tech
Blacksburg, Virginia, USA
surenthapa5803@gmail.com

Usman Naseem
Macquarie University
Sydney, New South Wales, Australia
usman.naseem@mq.edu.au

Abstract

Vaccine misinformation poses a persistent public health challenge, particularly in visual formats such as memes and infographics that combine text, imagery, and rhetorical cues. While textual misinformation has been widely studied, image-based vaccine misinformation remains comparatively underexplored due to the difficulty of interpreting multimodal signals at scale. In this work, we evaluate how effectively multimodal Large Vision-Language Models (LVLMs) can (i) directly classify vaccination stance from images and (ii) extract interpretable concept-level representations that support more reliable and transparent prediction. Using the VaxMeme dataset of 10,244 annotated images, we compare direct zero-shot LVLM inference against a hybrid framework in which classical machine learning models are trained on LVLM-extracted binary concept features. Our results show that grounding stance prediction in structured concept representations consistently outperforms direct LVLM classification, yielding accuracy improvements of approximately 10–17% while enabling explicit inspection of the visual and rhetorical cues driving model decisions. These findings highlight the value of concept-grounded, neuro-symbolic approaches for interpretable multimodal misinformation detection.

CCS Concepts

• **Computing methodologies** → **Information extraction**; **Artificial intelligence**; • **Social and professional topics** → *Computing / technology policy*; • **Applied computing** → *Document management and text processing*.

*Both authors contributed equally to this research and are joint first authors.
†Corresponding Author

WWW Companion '26, Dubai, United Arab Emirates

ACM ISBN 979-8-4007-2308-7/2026/04
https://doi.org/10.1145/3774905.3795453

Keywords

Vaccine misinformation, Visual memes, Multimodal misinformation detection, Vision-language models, Interpretable AI

ACM Reference Format:
Laxmi Thapa, Aryaman Jain*, Lakshmojee Koduru, Surabhi Adhikari, Junaid Rashid, Jungeun Kim, Surendrabikram Thapa, and Usman Naseem. 2026. Concept-Grounded Detection of Vaccine Misinformation in Multimodal Content Using Interpretable Vision-Language Models. In *Companion Proceedings of the ACM Web Conference 2026 (WWW Companion '26), April 13–17, 2026, Dubai, United Arab Emirates.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3774905.3795453

1 Introduction

Visual vaccine narratives have become a dominant form of online health discourse, often blending imagery, embedded text, and rhetorical framing to shape public perception [6, 15, 22, 25]. This shift challenges traditional misinformation detection pipelines and motivates multimodal approaches that can provide both reliable predictions and interpretable evidence for downstream decision-making.

1.1 The Crisis of Visual Misinformation and Existing Limitations

Vaccine misinformation remains a critical public health challenge, directly impacting immunization rates and global health security [10, 26]. While substantial academic effort has focused on detecting and mitigating textual misinformation, the rapid proliferation of visual content on social media, such as memes, infographics, and posters, presents a more complex and comparatively underexplored frontier [1, 4, 12]. These visual formats are particularly effective at bypassing text-based moderation systems because they communicate through nuanced multimodal cues, including humor, cultural symbolism, and emotional framing [2, 17].

Despite these risks, existing automated approaches to misinformation detection often struggle when applied to visual media [1, 5].

Traditional text-centric models are primarily optimized for linguistic patterns and fail to capture the interaction between imagery and embedded text that characterizes memes. Moreover, while recent Large Language Models (LLMs) and Vision-Language Models (LVLMs) demonstrate promising multimodal reasoning capabilities, their predictions are highly sensitive to prompt formulation, with small variations in phrasing leading to inconsistent stance classifications [16]. Finally, even when deep learning systems achieve strong performance, they often prioritize scalability over interpretability, limiting the ability of public health stakeholders to understand why specific visual content is flagged.

1.2 Research Objectives and Proposed Hybrid Framework

In this study, we investigate how effectively recent multimodal LVLMs capture visual and rhetorical cues in vaccine-related content, and whether concept-level representations can improve both interpretability and classification performance. Using the VaxMeme dataset [11], which consists of 10,244 manually annotated images spanning memes and mixed-media visuals, we evaluate four representative LVLMs: Qwen3-VL-30B-A3B[1] [20], Qwen3-8B[2], MiniCPM-V-2[3], and LLaVA-v1.5-7B[4]. For each image, we obtain two complementary outputs:

- **Direct Stance Classification:** The model predicts the vaccination stance (pro-vaccine, vaccine-critical, or neutral) based solely on its internal multimodal reasoning.
- **Structured Concept Extraction:** The model is prompted to identify 25 predefined binary concepts (e.g., *medical professional*, *questions safety*, *meme format*) in a valid JSON format, yielding an interpretable concept vector.

Using these extracted concept representations, we train classical machine learning classifiers to perform the final stance prediction. This hybrid design enables a systematic comparison between direct LVLM inference and concept-grounded classification. More broadly, the proposed framework allows us to assess the extent to which concept-level grounding improves predictive stability and interpretability, while also identifying the visual and rhetorical factors that most strongly influence vaccine stance. Together, these analyses provide a foundation for more transparent and reliable multimodal misinformation detection systems.

2 Related Work

We organize related work into several thematic areas, covering multimodal approaches to visual misinformation, concept bottleneck and post-hoc interpretability frameworks, prompt-based visual reasoning with large vision-language models, and neuro-symbolic methods for transparent social computing.

2.1 Visual Misinformation and Multimodal Analysis

The detection of misinformation has evolved from text-based approaches to increasingly complex multimodal frameworks. While early work focused primarily on linguistic patterns in social media posts, the rise of image-based propaganda, particularly internet memes, has necessitated the development of visual-semantic models [1]. Current state-of-the-art approaches typically employ end-to-end fusion strategies, in which image and text embeddings are concatenated and passed through dense classification layers [3, 19]. Although effective in terms of predictive accuracy, such models often lack transparency. In many cases, it is difficult to determine whether a prediction is driven by meaningful visual evidence or by spurious correlations, a limitation that is especially problematic in public health contexts where trust and accountability are critical.

2.2 Concept Bottleneck Models (CBMs)

To address the opacity of end-to-end neural networks, Koh et al. [7] introduced Concept Bottleneck Models (CBMs). Rather than mapping raw inputs (x) directly to predictions (y), CBMs first map inputs to a set of human-interpretable concepts (c), and then predict the target label solely based on these concepts ($x \rightarrow c \rightarrow y$). This architecture offers two key advantages: **interpretability**, as the model's reasoning is exposed through the concept layer, and **intervenability**, which allows human experts to correct mispredicted concepts to influence downstream decisions. However, standard CBMs require dense concept annotations during training, making them difficult to scale to rapidly evolving domains such as vaccine misinformation.

2.3 Post-hoc Concept Extraction via Multimodal Models

Recent work has sought to alleviate the annotation burden associated with CBMs. Yuksekgonul et al. [24] proposed Post-hoc Concept Bottleneck Models (PCBMs), which leverage pre-trained multimodal models such as CLIP to extract concepts without explicit concept-level supervision. By projecting images into a shared image–text embedding space, PCBMs define concepts using natural language descriptions. Our work extends this post-hoc paradigm by using generative Large Vision-Language Models (LVLMs) as the concept extraction mechanism. Unlike prior approaches that rely on embedding similarity scores, we prompt LVLMs to explicitly predict the presence of discrete rhetorical and semantic concepts, such as emotional framing or conspiratorial cues. This design enables the construction of a transparent concept bottleneck for vaccine stance detection while preserving the flexibility of open-vocabulary concept definitions.

2.4 Analysis of Extracted Concept Features

Beyond end-to-end classification, recent research has emphasized the value of analyzing the concept bottleneck layer itself as a structured semantic feature space. Oikarinen et al. [13] demonstrated that examining concept activation patterns allows researchers to quantify which attributes, such as institutional symbolism or conspiratorial framing, are most influential in driving model predictions. By treating extracted concepts as tabular features, classical interpretable models such as Logistic Regression and Decision Trees can be applied to assess feature importance. This hybrid methodology transforms opaque multimodal models into more transparent

[1] https://huggingface.co/Qwen/Qwen3-VL-30B-A3B-Instruct
[2] https://huggingface.co/Qwen/Qwen3-VL-8B-Instruct
[3] https://huggingface.co/openbmb/MiniCPM-V-2
[4] https://huggingface.co/liuhaotian/llava-v1.5-7b

systems, enabling systematic analysis of the visual and rhetorical factors associated with misinformation.

2.5 Prompt Engineering and Zero-Shot Visual Reasoning

With the transition from discriminative models to generative LVLMs, feature extraction has increasingly relied on prompt engineering. Prior work has shown that structured prompting strategies, including Chain-of-Thought prompting [18], can improve reasoning performance. In visual settings, such prompts encourage models to decompose complex images into constituent elements prior to classification. However, generative models remain susceptible to hallucination, describing objects or text that are not present in the input. As shown in HallusionBench [8], models such as LLaVA-v1.5-7B and Qwen exhibit sensitivity to prompt phrasing, highlighting the need for controlled and verifiable prompting strategies when LVLMs are used for automated data structuring.

2.6 Neuro-Symbolic Frameworks in Social Computing

Our approach aligns with broader trends in neuro-symbolic AI, which aim to combine the perceptual strengths of deep learning with the interpretability of symbolic reasoning. In computational social science, such hybrid systems typically separate perception from inference, using neural models to extract features from raw data and classical statistical models to reason about social phenomena. As noted by Yang et al. [21], disentangling feature extraction from prediction helps mitigate the risk of models exploiting spurious correlations, ensuring that decisions are grounded in semantically meaningful attributes.

2.7 Taxonomy of Concept Bottleneck Models

Concept Bottleneck Models have evolved substantially in recent years. Standard supervised CBMs [7] require datasets annotated with both labels y and dense concept vectors c, enabling joint learning of $x \to c$ and $c \to y$ mappings. While interpretable, this formulation suffers from an annotation bottleneck, as collecting fine-grained concept labels at scale is costly. Post-hoc CBMs [24] address this limitation by learning concept probes on top of fixed representations, though they may still suffer from concept leakage, where predictions implicitly rely on latent features.

More recent label-free and LLM-guided CBMs remove the need for manual concept annotation by using language models to propose candidate concepts. Methods such as LaBo [23] and OACE [13] generate concept descriptions from class labels and use vision–language models to estimate concept activations. These approaches are particularly well suited to dynamic domains such as misinformation, where new visual tropes and rhetorical strategies emerge frequently.

2.8 Recent Advances in Concept Utility and Optimization

Recent work has focused on improving the trade-off between interpretability and predictive performance in CBMs. Sparse concept bottlenecks [14] introduce regularization mechanisms that limit the number of active concepts, producing more concise and human-readable explanations. Hybrid and residual CBMs [9] incorporate latent feature pathways alongside explicit concepts, allowing models to capture visual nuances that are difficult to express linguistically. Finally, intervention-based methods leverage the explicit structure of CBMs to identify which concept corrections are most likely to influence predictions, enabling targeted human oversight in high-stakes settings such as misinformation detection.

3 Methodology

This section describes the proposed framework for evaluating vaccine stance in visual memes using multimodal Large Language Models (LLMs). We first formally define the vaccine stance classification task, then describe the dataset used in our experiments. Next, we introduce our LLM-based evaluation framework, including direct stance prediction and concept-level feature extraction. Finally, we outline the classical machine learning models and evaluation metrics used to assess performance.

3.1 Problem Formulation

Let $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{N}$ denote a dataset of N multimodal images, where x_i represents a visual meme or image containing both visual and textual cues, and $y_i \in \{\text{Pro}, \text{Critical}, \text{Neutral}\}$ denotes the ground-truth vaccination stance.

Our objective is to learn a function

$$f : \mathcal{X} \to \mathcal{Y}$$

that accurately predicts the vaccination stance from a given image.

3.2 Dataset

To evaluate the effectiveness of multimodal LLMs for vaccine stance detection, we use the **VaxMeme** dataset introduced by Naseem et al. [11]. The dataset consists of 10,244 manually annotated images collected from Twitter and was curated to address the lack of large-scale multimodal benchmarks for vaccine misinformation research.

The dataset exhibits substantial diversity in both visual format and rhetorical intent. As shown in Figure 1, the collection includes internet memes with overlaid text, public health infographics, screenshots of social media posts, excerpts from news reports, and handmade posters. These examples illustrate the wide range of visual strategies used to communicate vaccine-related narratives, ranging from emotionally charged and sarcastic memes to formal, informational content.

3.2.1 Dataset Labels. Each image in the VaxMeme dataset is annotated with one of three mutually exclusive vaccination stances: *Pro-vaccine*, *Vaccine-critical*, or *Neutral*. Pro-vaccine images promote vaccination or scientific evidence supporting vaccine safety and efficacy. Vaccine-critical images question vaccine safety, discourage immunization, or propagate conspiratorial narratives. Neutral images present descriptive or informational content without expressing a clear evaluative stance. This label taxonomy enables fine-grained analysis of multimodal vaccine discourse.

3.2.2 Dataset Statistics. Figure 2 illustrates the percentage distribution of stance labels in the dataset. Pro-vaccine content constitutes 38.9% of the dataset, followed by vaccine-critical content at 33.6%,

Figure 1: Representative examples from the VaxMeme dataset across different visual formats and vaccination stances. The examples illustrate the diversity of content, including text-based memes, infographics, social media screenshots, news-style visuals, and handmade posters, spanning pro-vaccine, vaccine-critical, and neutral narratives.

and neutral content at 27.5%. Although the dataset exhibits moderate class imbalance, all three stances are well represented. To mitigate potential performance bias arising from this imbalance, we adopt weighted evaluation metrics for precision, recall, and F1-score, which are described in detail in section 3.6.

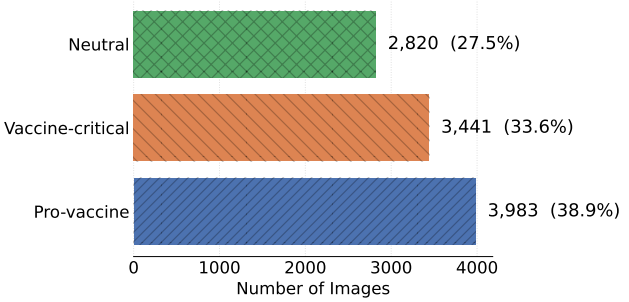

Figure 2: Percentage distribution of vaccination stance labels in the VaxMeme dataset.

3.3 LLM-Based Evaluation Framework

We evaluate multimodal Large Language Models (LLMs) under two complementary inference paradigms to assess their effectiveness in vaccine stance detection. Figure 3 provides an overview of the proposed concept-grounded evaluation framework, illustrating the parallel pathways of direct LVLM inference and concept-based classification. For each image in the dataset, we collect two distinct outputs from each LLM: (1) a direct stance prediction and (2) a structured concept-level representation. This dual-output framework enables a direct comparison between end-to-end LLM reasoning and concept-based classification.

- **Direct Stance Classification:** The LLM predicts a single vaccination stance label (*pro-vaccine*, *vaccine-critical*, or *neutral*) based solely on the input image, reflecting standard zero-shot multimodal inference.
- **Concept Feature Extraction:** The LLM is prompted to identify the presence or absence of a predefined set of semantic, rhetorical, and emotional concepts within the image. The output is a binary concept vector of length $K = 25$, returned in a strictly valid JSON format to enable deterministic parsing and downstream analysis.

By design, both outputs are generated from the same input image using identical model checkpoints. This allows us to isolate the effect of concept-level grounding on classification performance and interpretability. We compare the accuracy of direct LLM predictions against classical machine learning models trained on the extracted concept representations.

3.4 LLM Prompting Strategy

All LLM evaluations are conducted in a zero-shot setting, without any task-specific fine-tuning. We employ two distinct prompting strategies corresponding to the two inference paradigms described above.

3.4.1 Direct Stance Classification Prompting. For direct inference, the model is prompted to assign a single vaccination stance label to the input image. The model is instructed to respond with exactly one of the predefined labels and no additional text, reflecting a standard zero-shot multimodal classification setting. This configuration serves as a baseline for assessing the inherent multimodal reasoning capability of the LVLM.

Direct LVLM Classification. Formally, given an input image x_i, a Large Vision-Language Model (LVLM) predicts a stance label as:

$$\hat{y}_i = f_{\text{LVLM}}(x_i; \theta)$$

where $\hat{y}_i \in \{\text{pro-vaccine}, \text{vaccine-critical}, \text{neutral}\}$ and θ denotes the frozen model parameters. This formulation captures direct LVLM inference without any explicit concept-level grounding.

Zero-Shot Stance Classification Prompt

Given the above image, classify it as one of the following categories:

- pro-vaccine
- vaccine-critical
- neutral

Respond with only one label and no additional text.

3.4.2 Concept-Level Prompting. For concept extraction, the model is prompted to analyze the input image and return binary indicators

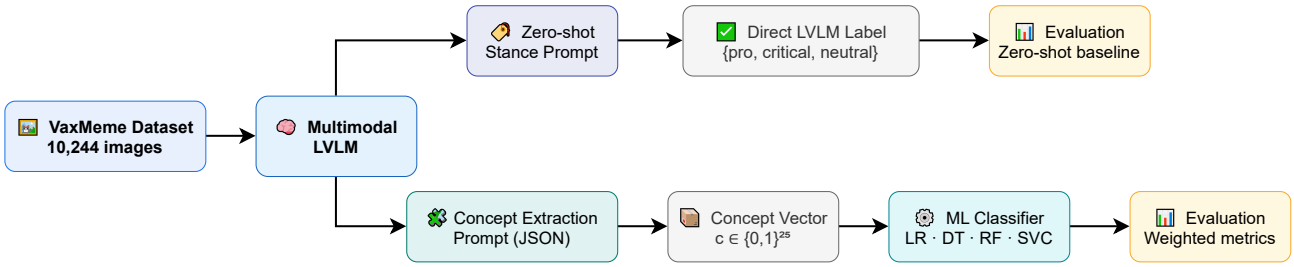

Figure 3: Overview of the proposed concept-grounded LVLM framework. Each image is processed by a multimodal LVLM to produce (i) a direct zero-shot stance prediction and (ii) a structured concept representation used by downstream machine learning classifiers. The symmetric design highlights the comparison between both approaches.

for a predefined set of 25 concepts capturing visual entities, rhetorical framing, emotional cues, and content format. The model is required to respond exclusively with a valid JSON object, ensuring consistent and machine-readable outputs. This structured prompting introduces an explicit semantic bottleneck between perception and prediction, enabling interpretable downstream modeling.

Concept Extraction via LVLM. Formally, given an input image x_i, the LVLM produces a concept activation vector:

$$\mathbf{c}_i = g_{\text{LVLM}}(x_i; \theta)$$

where $\mathbf{c}_i \in \{0,1\}^{25}$ represents the presence or absence of each predefined concept and θ denotes the frozen model parameters. Unlike direct stance prediction, this formulation exposes intermediate semantic representations that can be independently analyzed or used as features for downstream classifiers.

Concept Extraction Prompt

Analyze the given image and determine whether each of the following concepts is present. For each concept, output **1 if present** or **0 if absent**. Respond **only** with a valid JSON object and no additional text.

Concepts:

medical professional	virus depiction	fear
protester	fear symbol	conspiracy
child/family	promotes vaccination	freedom
questions safety	discourages vaccination	sarcasm
calls for freedom	politicized	hope
scientific evidence	anger	trust
humor	collective protection	meme format
official infographic	social media post	news report
handmade poster		

3.5 Machine Learning Classifiers for Stance Prediction

To transform the concept-level representations into final stance predictions, we train classical supervised machine learning models on the extracted binary concept vectors. Let $\mathbf{c}_i \in \{0,1\}^{25}$ denote the concept vector for image x_i, and let y_i be its corresponding stance label.

We evaluate four widely used classifiers to capture different inductive biases:

- **Logistic Regression:** A linear baseline used to assess the separability of stance labels in the concept space.
- **Decision Tree Classifier:** A non-linear model that provides explicit hierarchical decision rules and feature importance.
- **Random Forest Classifier:** An ensemble of decision trees used to improve generalization and reduce variance.
- **Support Vector Classifier (SVC):** A margin-based classifier used to identify optimal decision boundaries in the binary feature space.

For each classifier, we perform hyperparameter optimization using **GridSearchCV** with 5-fold cross-validation. Parameters such as the regularization strength (C) for Logistic Regression and SVC, and the maximum depth (`max_depth`) and number of estimators (`n_estimators`) for tree-based models are tuned to maximize validation performance. The dataset is split into training and testing sets using an **80/20 stratified split**, preserving the original class distribution across splits.

In addition to classification, Logistic Regression models are used for post-hoc interpretability by analyzing class-specific coefficient magnitudes, which serve as concept impact scores.

3.6 Evaluation Metrics

To assess the performance of both direct LVLM stance classification and downstream concept-based machine learning models, we employ four standard evaluation metrics. We evaluate both direct LVLM predictions and concept-based classifiers on the same fixed 20% held-out test split to ensure a fair comparison. As the task involves three classes (*Pro-vaccine, Vaccine-critical, Neutral*) with moderate class imbalance, we report weighted variants of precision, recall, and F1-score. Weighted averaging accounts for class prevalence by weighting each class-specific metric by its support in the ground truth distribution, ensuring a balanced and representative evaluation.

- **Accuracy:** The proportion of images whose predicted stance label matches the ground truth.
- **Weighted Precision:** The average precision across classes, weighted by class support. This metric reflects the reliability of the model when assigning stance labels.

- **Weighted Recall:** The average recall across classes, weighted by class support. It measures the model's ability to correctly identify instances of each stance.
- **Weighted F1-Score:** The harmonic mean of weighted precision and recall, providing a single robust measure that balances false positives and false negatives.

Formally, let $y_i \in \mathcal{Y}$ denote the ground-truth stance label for image x_i and $\hat{y}_i$ the corresponding prediction. Accuracy is defined as:

$$\text{Accuracy} = \frac{1}{N}\sum_{i=1}^{N}\mathbb{I}(\hat{y}_i = y_i),$$

where $\mathbb{I}(\cdot)$ is the indicator function and N is the number of test samples. For weighted precision, recall, and F1-score, we compute class-wise metrics and aggregate them using class support weights:

$$\text{Weighted Metric} = \sum_{k\in\mathcal{Y}} w_k \cdot \text{Metric}_k,$$

where $w_k = \frac{|\{i: y_i = k\}|}{N}$ denotes the proportion of samples belonging to class k.

These metrics are applied consistently across all experimental settings, including direct zero-shot LVLM inference and concept-based classification. This enables a fair comparison and allows us to quantify the performance gains achieved through structured concept extraction (e.g., the observed ≈10% improvement in accuracy).

4 Experimental Results

This section presents a comprehensive evaluation of the proposed concept-grounded framework. We first report zero-shot baseline performance of multimodal Large Vision-Language Models (LVLMs) on direct stance classification. We then analyze the semantic structure of LVLM-extracted concept features through post-hoc interpretability analysis. Finally, we compare direct LVLM inference against classical machine learning models trained on structured concept representations.

4.1 Zero-Shot LVLM Baselines

We begin by evaluating the performance of multimodal LVLMs under a direct zero-shot prompting setting, where each model is asked to assign a single vaccination stance label to an input image without any intermediate supervision or concept grounding. This experiment serves as a baseline for assessing the inherent multimodal reasoning capability of the models.

Table 1: Zero-shot stance classification performance of multimodal LVLMs using direct prompting.

Model	Accuracy	W-Prec.	W-Rec.	W-F1
LLaVA-v1.5-7B	0.4205	0.3945	0.4205	0.3437
MiniCPM-V-2	0.3886	0.4411	0.3886	0.3902
Qwen3-30B	0.5249	0.5599	0.5249	0.5328
Qwen3-8B	0.4949	0.5524	0.4949	0.5048

Table 1 reports the zero-shot stance classification performance across four LVLMs. Overall, accuracy ranges between approximately 39% and 52%, indicating that direct stance inference from multimodal inputs remains challenging. Among the evaluated models, Qwen3-30B achieves the highest accuracy and weighted F1-score, followed by Qwen3-8B, while LLaVA-v1.5-7B and MiniCPM-V-2 exhibit comparatively lower performance.

Despite moderate gains from larger model capacity, all LVLMs demonstrate substantial confusion across stance categories, particularly between vaccine-critical and neutral content. These results suggest that end-to-end LVLM reasoning alone is insufficient for reliably capturing the nuanced visual, rhetorical, and emotional cues present in vaccine-related memes, motivating the use of structured concept representations explored in subsequent experiments.

4.2 Concept Impact Analysis

Figures 4 and 5 present the concept impact analysis for features extracted using **Qwen3-8B** and **Qwen3-30B**, respectively. To identify the visual and rhetorical cues driving stance prediction, we analyze the class-specific coefficients of Logistic Regression models trained on the extracted concept representations. As a linear classifier, Logistic Regression provides direct interpretability, where the sign and magnitude of each coefficient indicate the direction and strength of a concept's contribution to a given stance.

Across both Qwen3 variants, the **Vaccine-critical** stance is strongly associated with concepts such as *questions safety*, *conspiracy*, and *calls for freedom*. These features reflect well-established anti-vaccine narratives centered on distrust of institutional authority and appeals to individual autonomy. Notably, Qwen3-30B assigns substantially higher weights to affective concepts such as *anger* and *fear symbol*, suggesting that the larger model captures emotional framing more effectively than Qwen3-8B.

In contrast, the **Pro-vaccine** stance is primarily driven by concepts including *promotes vaccination*, *medical professional*, and *hope*, highlighting the role of scientific authority and positive health messaging in pro-vaccine discourse.

For the **Neutral** stance, concepts such as *collective protection* and *news report* receive the highest impact scores. This indicates that images adopting a formal, informational presentation style without strong emotional or rhetorical cues are more likely to be interpreted as non-ideological.

Overall, this analysis demonstrates that concept representations extracted by Qwen3-based LVLMs form a coherent and interpretable feature space. The clear separation of high-impact concepts across stances confirms that the proposed concept schema effectively captures the dominant rhetorical and emotional strategies underlying vaccine-related visual misinformation.

4.3 Concept-Based Classification Performance

We next evaluate the effectiveness of using LVLM-extracted concept representations for downstream vaccine stance classification. Using the binary concept vectors as input features, we train four classical machine learning models: Logistic Regression, Decision Tree, Support Vector Classifier (SVC), and Random Forest. Performance is compared against direct zero-shot LVLM inference to assess the benefits of concept grounding.

Across all evaluated LVLMs and classifiers, concept-based models consistently outperform direct LVLM predictions. Among the tested algorithms, Random Forest achieves the strongest overall

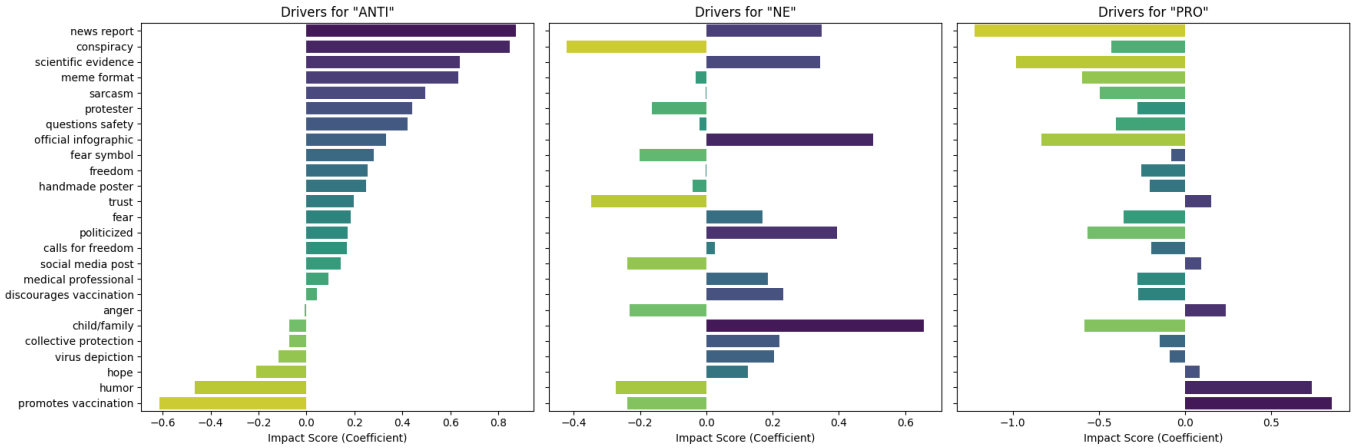

Figure 4: Concept impact analysis for Qwen3-8B. Bars indicate class-specific Logistic Regression coefficients for LVLM-extracted concept features, highlighting the most influential drivers for each vaccination stance.

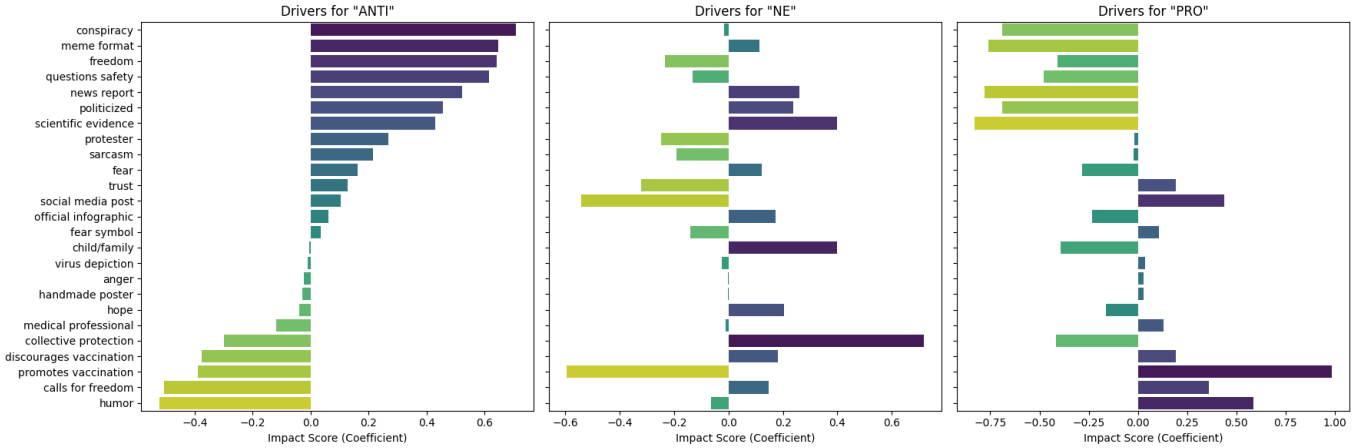

Figure 5: Concept impact analysis for Qwen3-30B. Compared to Qwen3-8B, the larger model exhibits stronger sensitivity to affective and rhetorical concepts such as anger and fear-related cues.

Table 2: Concept-based stance classification performance using Random Forest classifiers trained on LVLM-extracted features.

LVLM	Accuracy	W-Prec.	W-Rec.	W-F1
LLaVA-v1.5-7B	0.4835	0.5046	0.4835	0.4340
MiniCPM-V-2	0.5612	0.5527	0.5612	0.5392
Qwen3-30B	0.6458	0.6431	0.6458	0.6423
Qwen3-8B	0.6159	0.6067	0.6159	0.5971

performance, indicating that ensemble-based non-linear models are particularly effective at capturing interactions between semantic concepts.

Table 2 presents the performance of Random Forest classifiers trained on concept features extracted by different LVLMs. Compared to zero-shot baselines, accuracy improves by approximately 5–15% across models, with corresponding gains in weighted precision, recall, and F1-score. Concept features extracted by Qwen3-30B yield the highest performance, achieving an accuracy of 0.646 and a weighted F1-score of 0.642.

Table 3: Concept-based stance classification performance using Decision Tree trained on LVLM-extracted features.

LVLM	Accuracy	W-Prec.	W-Rec.	W-F1
LLaVA-v1.5-7B	0.4727	0.4917	0.4727	0.4222
MiniCPM-V-2	0.5317	0.5184	0.5317	0.5139
Qwen3-30B	0.6349	0.6310	0.6349	0.6319
Qwen3-8B	0.6179	0.6132	0.6179	0.6134

Tables 3, 4, and 5 report results for Decision Tree, Logistic Regression, and SVC classifiers, respectively. While these models generally underperform Random Forests, they nonetheless demonstrate consistent improvements over direct LVLM inference, confirming the

robustness of the concept-based approach across different classifier families.

Overall, these results demonstrate that grounding stance prediction in structured, interpretable concept representations leads to substantial and consistent performance gains over direct LVLM classification. This finding supports the central premise of the proposed framework: decoupling perception and prediction improves both predictive accuracy and interpretability.

Table 4: Concept-based stance classification performance using Logistic Regression model trained on LVLM-extracted features.

LVLM	Accuracy	W-Prec.	W-Rec.	W-F1
LLaVA-v1.5-7B	0.4771	0.5322	0.4771	0.4137
MiniCPM-V-2	0.5189	0.4925	0.5189	0.4707
Qwen3-30B	0.6254	0.6217	0.6254	0.6168
Qwen3-8B	0.5910	0.5798	0.5910	0.5708

Table 5: Concept-based stance classification performance using SVC trained on LVLM-extracted features.

LVLM	Accuracy	W-Prec.	W-Rec.	W-F1
LLaVA-v1.5-7B	0.4697	0.5148	0.4697	0.4059
MiniCPM-V-2	0.5166	0.4897	0.5166	0.4602
Qwen3-30B	0.6175	0.6132	0.6175	0.6102
Qwen3-8B	0.5988	0.5999	0.5988	0.5944

5 Discussion and Future Work

5.1 Interpretability Benefits of Concept-Grounded Models

Beyond improvements in predictive performance, a key contribution of this work is the shift from opaque end-to-end classification toward a concept-grounded framework that enables transparent decision-making. In direct LVLM prompting, a model may correctly assign a vaccine-critical label, but the reasoning process remains implicit and difficult to verify. Generated explanations in such settings are often post-hoc and may not reflect the actual signals used for prediction.

By decomposing each image into a structured vector of binary concepts, the proposed framework introduces an explicit semantic bottleneck between perception and prediction. This design forces the system to surface intermediate evidence—such as the presence of conspiratorial framing or emotionally charged symbols—prior to final stance classification. As a result, misclassifications can be inspected and traced back to specific concept activations, enabling more targeted analysis and facilitating human-in-the-loop review. This level of transparency is particularly important in public health applications, where trust and accountability are essential.

5.2 Limitations and Future Directions

Despite its advantages, the proposed approach has several limitations that motivate future research.

5.2.1 Fixed Concept Vocabulary. The current framework relies on a predefined set of 25 concepts, which assumes prior knowledge of salient visual and rhetorical cues in vaccine discourse. However, misinformation narratives evolve rapidly, and new visual tropes may emerge over time. Future work could explore adaptive or dynamic concept discovery mechanisms, in which LLMs periodically propose and validate new concepts based on emerging content.

5.2.2 Loss of Fine-Grained Visual Information. Binary concept representations necessarily compress rich visual signals. Subtle cues related to visual style, composition, or tone may be lost during discretization. Incorporating residual or hybrid representations that combine explicit concepts with low-dimensional latent features may help capture such nuances while preserving interpretability.

5.2.3 Contextual and Cultural Ambiguity. The current concept schema primarily captures literal visual and textual indicators and may struggle with culturally specific references, irony, or sarcasm. Future extensions should incorporate relational or reasoning-based prompts that explicitly model interactions between text and imagery for robust handling of contextual and ironic content.

6 Conclusion

Visual misinformation poses a growing challenge for public health, particularly due to the semantic complexity of multimodal content such as memes, infographics, and hybrid image–text artifacts. Traditional text-centric NLP systems are not well suited to this setting. In this work, we evaluated the effectiveness of state-of-the-art multimodal Large Vision-Language Models (LVLMs) for vaccine stance detection and examined their role beyond direct end-to-end classification. Our experiments show that direct zero-shot LVLM inference yields unstable performance, with accuracy typically ranging between 40% and 50%. However, when LVLMs are repurposed as semantic feature extractors, their capabilities become substantially more reliable. Grounding stance prediction in a structured space of interpretable visual and rhetorical concepts leads to consistent performance improvements of approximately **10–17%** across multiple model architectures and evaluation metrics. More broadly, this study demonstrates the practical value of concept-grounded, neuro-symbolic pipelines for multimodal misinformation analysis. By decoupling perception from prediction, the proposed framework combines the representational strength of generative LVLMs with the robustness and transparency of classical machine learning models. This design not only improves predictive accuracy but also enables systematic inspection of the visual and rhetorical cues driving model decisions. Such properties are critical for real-world misinformation monitoring systems, where interpretability, accountability, and human oversight are essential.

7 Acknowledgements

This work was partly supported by the Institute of Information & Communications Technology Planning & Evaluation(IITP)-ITRC (Information Technology Research Center) grant funded by the Korea government (MSIT) (IITP-2026-RS-2024-00438335) and by the Technology Innovation Program funded By the Ministry of Trade, Industry & Energy(MOTIE)(No.20022899).

References

[1] Sara Abdali, Sina Shaham, and Bhaskar Krishnamachari. 2024. Multi-modal misinformation detection: Approaches, challenges and opportunities. *Comput. Surveys* 57, 3 (2024), 1–29.

[2] Aashish Bhandari, Siddhant B Shah, Surendrabikram Thapa, Usman Naseem, and Mehwish Nasim. 2023. Crisishatemm: Multimodal analysis of directed and undirected hate speech in text-embedded images from russia-ukraine conflict. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 1994–2003.

[3] Anusha Chhabra and Dinesh Kumar Vishwakarma. 2023. A literature survey on multimodal and multilingual automatic hate speech identification. *Multimedia Systems* 29, 3 (2023), 1203–1230.

[4] Carmela Comito, Luciano Caroprese, and Ester Zumpano. 2023. Multimodal fake news detection on social media: a survey of deep learning techniques. *Social Network Analysis and Mining* 13, 1 (2023), 101.

[5] Dimitar Dimitrov, Bishr Bin Ali, Shaden Shaar, Firoj Alam, Fabrizio Silvestri, Hamed Firooz, Preslav Nakov, and Giovanni Da San Martino. 2021. Detecting propaganda techniques in memes. In *Proceedings of the 59th annual meeting of the association for computational linguistics and the 11th international joint conference on natural language processing (volume 1: long papers)*. 6603–6617.

[6] Brian Hughes, Cynthia Miller-Idriss, Rachael Piltch-Loeb, Beth Goldberg, Kesa White, Meili Criezis, and Elena Savoia. 2021. Development of a codebook of online anti-vaccination rhetoric to manage COVID-19 vaccine misinformation. *International journal of environmental research and public health* 18, 14 (2021), 7556.

[7] Pang Wei Koh, Thao Nguyen, Yew Siang Tang, Stephen Mussmann, Emma Pierson, Been Kim, and Percy Liang. 2020. Concept Bottleneck Models. In *International Conference on Machine Learning (ICML)*. PMLR, 5338–5348.

[8] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2024. HallusionBench: You See What You Think? Or You Think What You See? An Image-Context Reasoning Benchmark Challenging for GPT-4V(ision), LLaVA-1.5, and Gemini. In *CVPR*.

[9] Yang Liu, Tianwei Zhang, and Shi Gu. 2025. Hybrid Concept Bottleneck Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 20179–20189.

[10] Sahil Loomba, Alexandre De Figueiredo, Simon J Piatek, Kristen De Graaf, and Heidi J Larson. 2021. Measuring the impact of COVID-19 vaccine misinformation on vaccination intent in the UK and USA. *Nature human behaviour* 5, 3 (2021), 337–348.

[11] Usman Naseem, Jinman Kim, Matloob Khushi, and Adam G. Dunn. 2023. A Multimodal Framework for the Identification of Vaccine Critical Memes on Twitter. In *International Conference on Web Search and Data Mining (WSDM)*.

[12] Usman Naseem, Imran Razzak, Katarzyna Musial, and Muhammad Imran. 2020. Transformer based deep intelligent contextual embedding for twitter sentiment analysis. *Future Generation Computer Systems* 113 (2020), 58–69.

[13] Tuomas Oikarinen, Subhro Das, Lam M Nguyen, and Tsui-Wei Weng. 2023. Label-free Concept Bottleneck Models. In *International Conference on Learning Representations (ICLR)*.

[14] Andrei Semenov, Vladimir Ivanov, Aleksandr Beznosikov, and Alexander Gasnikov. 2024. Sparse Concept Bottleneck Models: Gumbel Tricks in Contrastive Learning. *arXiv preprint arXiv:2404.03323* (2024).

[15] Elizabeth A Shanahan, Rob A DeLeo, Elizabeth A Albright, Meng Li, Elizabeth A Koebele, Kristin Taylor, Deserai Anderson Crow, Katherine L Dickinson, Honey Minkowitz, Thomas A Birkland, et al. 2023. Visual policy narrative messaging improves COVID-19 vaccine uptake. *PNAS nexus* 2, 4 (2023), pgad080.

[16] Surendrabikram Thapa, Shuvam Shiwakoti, Siddhant Bikram Shah, Surabhi Adhikari, Hariram Veeramani, Mehwish Nasim, and Usman Naseem. 2025. Large language models (llm) in computational social science: prospects, current state, and challenges. *Social Network Analysis and Mining* 15, 1 (2025), 1–30.

[17] Surendrabikram Thapa, Hariram Veeramani, Imran Razzak, Roy Ka-Wei Lee, and Usman Naseem. 2025. Cross Platform MultiModal Retrieval Augmented Distillation for Code-Switched Content Understanding. In *Companion Proceedings of the ACM on Web Conference 2025*. 2042–2051.

[18] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Ed Chi, Quoc Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems (NeurIPS)*.

[19] Qingzheng Xu, Heming Du, Szymon Łukasik, Tianqing Zhu, Sen Wang, and Xin Yu. 2025. MDAM3: A Misinformation Detection and Analysis Framework for Multitype Multimodal Media. In *Proceedings of the ACM on Web Conference 2025*. 5285–5296.

[20] An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388* (2025).

[21] Jing Yang et al. 2023. Interpretable Multimodal Misinformation Detection with Logic Reasoning. *Findings of the Association for Computational Linguistics (ACL)* (2023).

[22] Shiyu Yang, Dominique Brossard, Nan Li, and Leonardo Barolo Gargiulo. 2024. Bridging gaps in COVID-19 vaccine knowledge: Effects of multimodal narratives on message elaboration and recall across science literacy levels. *Clinical Epidemiology and Global Health* 28 (2024), 101681.

[23] Yue Yang, Artemis Panagopoulou, Shenghao Zhou, Daniel Jin, Chris Callison-Burch, and Mark Yatskar. 2023. Language in a Bottle: Language Model Guided Concept Bottlenecks for Interpretable Image Classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 19187–19197.

[24] Mert Yuksekgonul, Maggie Wang, and James Zou. 2023. Post-hoc Concept Bottleneck Models. In *International Conference on Learning Representations (ICLR)*.

[25] Agata Zdun-Ryżewska, Wiktoria Grabowska, Katarzyna Milska-Musa, Natalia Nadrowska, and Magdalena Błażek. 2025. Mapping Vaccine Narratives and Emotional Framing in Ideologically Divergent Social Media Communities During COVID-19. *language* 20 (2025), 21.

[26] Sihong Zhao, Simeng Hu, Xiaoyu Zhou, Suhang Song, Qian Wang, Hongqiu Zheng, Ying Zhang, Zhiyuan Hou, et al. 2023. The prevalence, features, influencing factors, and solutions for COVID-19 vaccine misinformation: systematic review. *JMIR public health and surveillance* 9, 1 (2023), e40201.