LilyMeme@EEUCA 2026: Multimodal Vaccine Meme Stance Detection with Task-Adapted MemeCLIP and Complementary Ensembling

Yixuan Li[1], Xiaolong Yin[2], Yang Yang[1]*
[1]Nanjing University of Science and Technology
[2]Nanjing University
liyixuan25@njust.edu.cn, yinxl@lamda.nju.edu.cn, yyang@njust.edu.cn

Abstract

Memes have emerged as a prominent medium for conveying public sentiment on sensitive health topics such as vaccination. Unlike conventional multimodal tasks, memes feature implicit stances, sarcastic nuances, and complex cross-modal interactions, posing significant challenges for accurate stance detection. This paper presents our approach for the VaxMeme Shared Task @EEUCA 2026, which aims to classify vaccine-related memes into three distinct classes: Vaccine-critical, Neutral, and Pro-vaccine. Building upon MemeCLIP, we systematically enhance our framework via task-specific adaptation, lightweight cross-modal fusion, noise-aware training, LLM-assisted semantic augmentation, and inference-stage optimization, ultimately ensembling multiple complementary variants for final predictions. Our ensemble method achieves a Macro-F1 score of 0.8494 on the official test set, securing first place and demonstrating the critical efficacy of noise-aware training and late-stage ensembling for robust stance identification.

1 Introduction

With the development of social media, memes have become an important medium for expressing viewpoints on public issues, especially on sensitive topics such as vaccination, public health, and misinformation. Unlike ordinary text or images, memes often rely jointly on images, short text, embedded image text, sarcastic rhetoric, and symbolic visual elements to convey opinions. Therefore, automatically identifying their stance is more difficult than general classification tasks. The Shared Task on Multimodal Vaccine Critical Meme Detection (VaxMeme) @EEUCA 2026 was proposed precisely around this challenge (Thapa et al., 2026b; Hürriyetoğlu et al., 2026). This task focuses on fine-grained stance understanding in public health scenarios and provides a valuable benchmark for studying multimodal reasoning under noise, indirect expression, and social context.

Figure 1: An example from the shared task dataset labeled as Vaccine critical.

This shared task is challenging in multiple aspects. First, memes usually contain heterogeneous information sources, which may be complementary, partially redundant, or even semantically conflicting. Second, vaccine-related memes often rely on sarcasm, exaggeration, and implicit rhetoric, making literal interpretation of the text often unreliable. Third, some samples contain only weak textual evidence, noisy OCR, or highly compressed symbolic visual content, so the model must rely on subtle multimodal cues rather than explicit sentiment words to determine stance (Naseem et al., 2023; Thapa et al., 2026b).

To address these difficulties, we take MemeCLIP (Shah et al., 2024) as the core framework and design multiple enhancement methods for the VaxMeme shared task. We first reorganize the input for VaxMeme, introduce explicit missing-text markers, replace the original simple fusion strategy with lightweight token-level cross-modal interaction, and strengthen training with stratified

*Corresponding author.

Proceedings of the 9th Workshop on Event Extraction and Understanding: Challenges and Applications, pages 208–215
July 3, 2026 ©2026 Association for Computational Linguistics

cross-validation, class weighting, label smoothing, and cosine learning rate scheduling. On this basis, we further explore multiple complementary enhancement methods, including noise-aware weighted training, an LLM-assisted semantic description branch, a three-branch architecture that explicitly distinguishes post text from OCR text, and inference-stage enhancement that combines test-time augmentation with retrieval priors. Our final submission does not rely on any single model, but is instead obtained by ensembling multiple complementary variants.

2 Related Work

In recent years, multimodal meme understanding has evolved from coarse-grained harmful content detection to fine-grained pragmatic and stance analysis (Guan et al., 2025). Early research predominantly focused on identifying hate speech, offensive content, and humor; however, the proliferation of novel datasets and advanced vision-language models has shifted attention toward more complex semantic properties embedded within memes, including targets, stances, and implicit contexts.

Recent studies (Liang et al., 2024; Yang et al., 2024) have systematically explored text-image stance detection by curating diverse social media datasets and introducing target-aware cross-modal prompting strategies. Furthermore, the introduction of the PrideMM dataset and the associated MemeCLIP framework (Shah et al., 2024) transitioned meme analysis from isolated harmful content detection to a comprehensive multi-task paradigm encompassing hate, target, stance, and humor recognition. Subsequently, the CASE 2025 Shared Task formalized multimodal stance recognition in meme scenarios by establishing it as an independent evaluation track (Thapa et al., 2025). Collectively, these advancements highlight that the core challenge of meme stance recognition extends beyond naive image-text fusion; it necessitates the intricate modeling of multimodal synergies, contradictions, and the underlying socio-cultural contexts (Yu et al., 2026; Jiang et al., 2025).

In comparison, research on stance classification for vaccine-related memes remains relatively limited. MMCoVaR provides a multimodal dataset for COVID-19 vaccines, covering both news and tweets, for misinformation- and credibility-related classification tasks (Chen et al., 2021). (Naseem et al., 2023) were the first to systematically introduce the VaxMeme task and dataset, collecting a large-scale manually annotated set of vaccine-related memes and designing a multimodal framework that combines global and local representations for identifying vaccine-critical memes. More recent work has also started to extend multimodal vaccine-content analysis from coarse-grained classification toward more interpretability-oriented directions, and the EEUCA 2026 Shared Task can be seen as a natural extension of earlier multimodal meme stance research into the public-health domain (Thapa et al., 2026a,b).

3 Task and Dataset

3.1 Task Definition

The Shared Task on Multimodal Vaccine Critical Meme Detection (VaxMeme) @EEUCA 2026 is a three-class multimodal stance identification task (Thapa et al., 2026b; Hürriyetoğlu et al., 2026). For each meme with a unique identifier `index`, the model is required to determine its stance toward vaccination, namely Vaccine critical, Neutral, or Pro-vaccine. The shared task adopts Macro-F1 as the primary evaluation metric, which means that the model must not only achieve strong overall classification performance, but also maintain as balanced recognition performance as possible across the three classes. The challenge of this task mainly comes from the complexity of meme expression. On the one hand, the stance of a meme is often not directly expressed through a single modality, but is jointly conveyed through the interaction between image and text. On the other hand, sarcasm, exaggeration, metaphor, and image–text incongruity are very common in this type of data, making it difficult to fully model the true semantics by relying only on visual features or only on textual features.

3.2 Dataset

This study utilizes the official dataset curated for the EEUCA 2026 Shared Task on Multimodal Vaccine Critical Meme Detection. This corpus is primarily derived from the VaxMeme dataset alongside associated data collection initiatives (Thapa et al., 2026b; Naseem et al., 2023; Bhandari et al., 2023; Thapa et al., 2026a). Originally introduced by (Naseem et al., 2023) for multimodal vaccine-critical meme identification, VaxMeme comprises over 10,000 English meme samples sourced from Twitter. These samples encapsulate both visual and textual modalities, where the embedded text within

Table 1: Class distribution of image samples in the shared task dataset.

Dataset	Vaccine critical	Neutral	Pro-vaccine	Total
Train	2535	2461	3199	8195
Val	308	327	389	1024
Test	314	316	395	1025

the images is automatically extracted via optical character recognition (OCR) (Naseem et al., 2023). Following the official shared task configuration, the dataset is partitioned into training, validation, and test sets, comprising 8,195, 1,024, and 1,025 instances, respectively. Each instance consists of a meme image, the associated optical character recognition output (`image_text`), and the corresponding social media post (`post_text`). The classification schema encompasses three stance categories: Pro-vaccine, Vaccine-critical, and Neutral, with the detailed class distribution summarized in Table 1.

4 Method

4.1 MemeCLIP Baseline

Our work is built upon the MemeCLIP baseline (Shah et al., 2024). MemeCLIP is a CLIP-based framework for text-embedded meme classification, whose core idea is to preserve the pre-trained knowledge of CLIP while only performing lightweight adaptation on the downstream classification modules. Specifically, the original implementation adopts CLIP ViT-L/14 as the vision–language backbone and freezes its parameters (Radford et al., 2021); for each meme, the model first extracts image features and text features separately, and then maps them into a shared feature space through independent linear projection layers. Subsequently, both the image branch and the text branch pass through lightweight Adapters and are residually mixed with the original projected features. The original cross-modal fusion is implemented as element-wise multiplication between the image and text representations, and the final classification prediction is produced by a lightweight feed-forward layer together with a cosine classifier (Shah et al., 2024).

4.2 MemeCLIP for VaxMeme

MemeCLIP provides a starting point for text-embedded meme understanding, but it was not directly designed for VaxMeme. Therefore, we adapt MemeCLIP based on the requirements of the shared task.

First, we reorganize the original Pride-based meme classification setting into a three-class task for VaxMeme, and rebuild the data processing pipeline. We construct a unified metadata file and split the training/validation data using five-fold StratifiedKFold. Second, in terms of input construction, we no longer follow a single text field, but explicitly integrate two text sources: the post-level text `post_text` and the embedded text in the image, `image_text`, and organize them as a structured template

[POST] `post_text` [IMG] `image_text`.

We further introduce explicit missing markers `[NO_POST]` and `[NO_OCR]` to distinguish "missing text" from ordinary empty input. To improve robustness, we also apply lightweight text dropout during training, so as to reduce the model's over-reliance on any single textual signal.

To enhance the multimodal fusion mechanism, we upgrade the naive element-wise operation originally employed in MemeCLIP to a lightweight cross-modal Transformer. The input sequence to this module comprises three distinct entities: a parameterized `[CLS]` token, an image token, and a text token. Following modality-specific embedding and self-attention-driven interaction, the ultimate fused representation aggregates the updated `[CLS]` state with the cross-modal interaction terms derived from the visual and textual tokens. This architectural refinement substantially augments the capacity of the network to capture fine-grained, token-level cross-modal alignments.

4.3 Enhancement Methods

Noise-Aware Weighted Training. We observe that the vaccine meme data may contain a certain proportion of highly ambiguous samples or suspiciously mislabeled samples, and therefore further explore a noise-aware weighted training method. In the offline stage, this method constructs nearest-neighbor consistency analysis based on image features, text features, and fused features, respectively, and computes the agreement between each sample and the labels of its neighbors, thereby deriving a noise score and conflict count for each sample. Subsequently, instead of directly removing suspected noisy samples, we convert them into different training weights `sample_weight`, so as to reduce the contribution of suspicious samples during training.

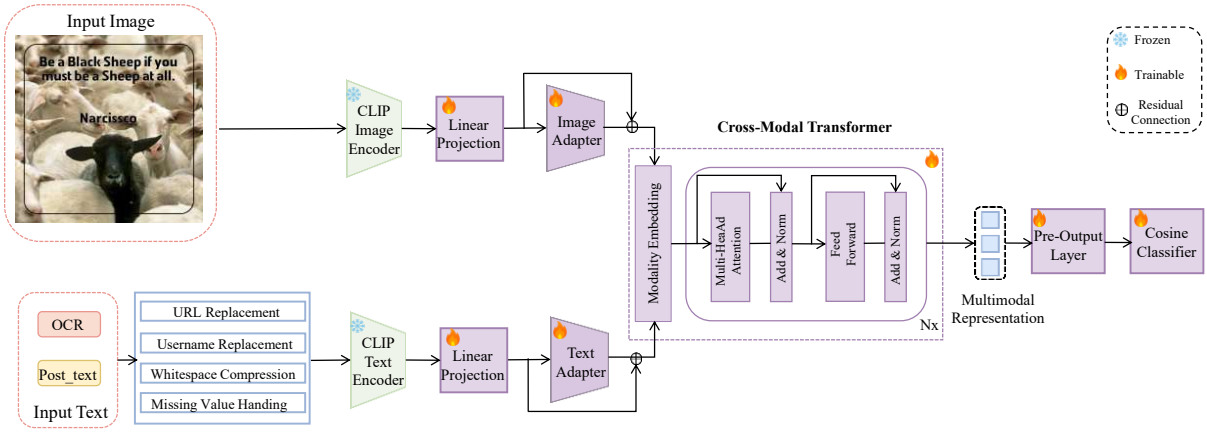

Figure 2: Simplified Architecture of the Adapted MemeCLIP Framework.

LLM-Assisted Multimodal Fusion. We attempt to use large language models to provide additional semantic supplementation. Specifically, we introduce an LLM-assisted multimodal fusion method by adding an auxiliary description branch, `llm_desc`. These descriptions are generated by Qwen2.5-VL-7B-Instruct for memes with poor OCR quality (Bai et al., 2025). Through this selective generation strategy, we aim to examine whether neutral visual supplementary descriptions generated by an LLM can provide additional benefit for vaccine meme stance classification when OCR evidence is insufficient.

During generation, we use prompts that require the model to output neutral and factual supplementary descriptions, focusing on visible persons, objects, actions, symbols, and layout information, while explicitly prohibiting the inference of hidden intent or direct prediction of stance categories. Structurally, we extend the fusion tokens from three to four, namely `[CLS]`, image, text, and description. The final fused representation combines the updated `[CLS]` representation together with three types of interaction terms: image–text, image–description, and text–description. We additionally apply a small-probability dropout to the description branch during training to avoid over-reliance on this auxiliary input.

Three-Branch Multimodal Fusion. We aim to distinguish text information from different sources in a more fine-grained manner. In the standard adapted version, `post_text` and `image_text` are concatenated as a single text input; in this variant, however, we explicitly model them separately, forming a three-branch multimodal fusion architecture: an image branch, a post-text branch, and an OCR-text branch. The motivation for this design is that these two types of text naturally play different semantic roles: `post_text` often serves as contextual supplementation, whereas `image_text` is more likely to provide the most direct stance evidence inside the image.

In this architecture, the image, post text, and OCR text each extract frozen CLIP features, and are then adapted into a shared hidden space through their own projection layers and Adapters(Radford et al., 2021). After residual mixing and normalization, the three branches, together with a learnable `[CLS]` token, are fed into a lightweight cross-modal Transformer. The final fused representation is composed of the updated `[CLS]` representation and the three pairwise interaction terms. Compared with the standard adapted version, this design preserves the branch-specific characteristics of different text sources while modeling their relations through a unified token-level interaction mechanism.

Inference-Stage Enhancement. Beyond architectural optimizations and training-level enhancements, we introduce inference-stage refinement by integrating lightweight test-time augmentation (TTA) with retrieval-augmented prediction. Specifically for TTA, we generate a set of conservative augmented views for each meme image, encompassing the standard CLIP preprocessing pipeline and resizing along the shorter edge followed by center cropping (Radford et al., 2021). To implement retrieval augmentation, we construct a global feature bank comprising all training samples, where the representation of each instance is derived by averaging its fused features extracted across multiple cross-validation folds. During inference, we compute the averaged fused feature for an incoming

test sample and subsequently retrieve its nearest neighbors from the feature bank utilizing cosine similarity. Utilizing the ground-truth labels of the top-k neighbors alongside their temperature-scaled similarity weights, we formulate a k-nearest neighbor (knn) probability distribution, denoted as p_{knn}. Ultimately, the parametric model prediction p_{model} is interpolated with the non-parametric retrieval prior p_{knn} to yield the final probability distribution:

$$p_{\text{final}} = \alpha p_{\text{model}} + (1 - \alpha) p_{\text{knn}}$$

This formulation effectively harnesses both the generalization capabilities inherent in the parametric model and the robust, instance-level evidence provided by non-parametric nearest neighbors.

Final Ensemble System. Our final submission relies on an ensemble framework comprising multiple complementary model variants rather than a single monolithic architecture. Specifically, we integrate models derived from different cross-validation folds alongside variants diversified across visual backbones, robustness optimization, text-source modeling, auxiliary semantics, and inference-stage enhancements. Within this integrated framework, the task-adapted MemeCLIP (Shah et al., 2024) serves as a stable, lightweight baseline, while more advanced backbones extract superior foundational representations. Furthermore, noise-aware training mitigates the interference from ambiguous instances, the three-branch architecture explicitly disentangles heterogeneous text sources, and inference-stage augmentations introduce nearest-neighbor priors to bolster generalization. Consequently, this integrated system exemplifies a highly pragmatic methodology for tasks, maximizing both predictive accuracy and systemic stability by fully exploiting the synergistic complementarity among distinct configurations.

5 Experimental Setup

Data Preprocessing. We implement a standardized preprocessing pipeline across all textual modalities to ensure semantic consistency. Specifically, within the `post_text` and `image_text` fields, we systematically resolve null values and normalize noisy artifacts, including URLs, user mentions, and redundant whitespace. To explicitly encode the absence of specific modalities without losing structural information, missing text entries are substituted with predefined special tokens, namely `[NO_POST]` and `[NO_OCR]`.

Table 2: Results of different model variants on the official shared-task test set, where T stands for the Three-branch Multimodal Fusion method, L stands for the LLM-assisted Multimodal Fusion method, W stands for the Noise-aware Weighted Training method, and I stands for the Inference-stage Enhancement method.

Model	Macro-F1	Accuracy	Precision	Recall
CLIP ViT-L/14	0.8145	0.8166	0.8191	0.8154
EVA-CLIP	0.8170	0.8195	0.8179	0.8182
EVA-CLIP+T	0.8159	0.8185	0.8173	0.8185
EVA-CLIP+L	0.7980	0.8000	0.8035	0.7984
EVA-CLIP+W	0.8239	0.8263	0.8243	0.8256
EVA-CLIP+W+I	0.8355	0.8380	0.8361	0.8377
ENSEMBLE MODEL	**0.8494**	**0.8517**	**0.8494**	**0.8517**

Model Settings and Hyperparameters. Our primary architecture is constructed upon a frozen vision-language backbone. The standard configuration employs CLIP ViT-L/14 (Radford et al., 2021), whereas the more advanced variant utilizes EVA02-L-14, implemented via OpenCLIP (Sun et al., 2023). Across the primary adapted model and most subsequent variants, the core hyperparameters are uniformly set as follows: an input image resolution of 224, a batch size of 16, a residual mixing ratio of 0.2, and a cosine classifier scaling factor of 30. The lightweight cross-modal Transformer comprises 2 layers and 8 attention heads, with dropout rates configured as [0.10, 0.30, 0.20] across its internal components. Optimization is performed using AdamW with a learning rate of 3×10^{-5}, a weight decay of 5×10^{-4}, and a maximum of 10 training epochs. All experiments are executed on a NVIDIA GeForce RTX 3090 GPU.

For the LLM-enhanced variant, we incorporate Qwen2.5-VL-7B-Instruct (Bai et al., 2025) to generate auxiliary textual descriptions. To mitigate the introduction of superfluous noise, this generation process is strictly triggered only for samples exhibiting poor OCR quality. The generation prompt is carefully crafted to elicit neutral and factual visual supplements, explicitly prohibiting the model from inferring hidden intents or directly predicting stance categories. Consequently, this descriptive branch functions exclusively as a supplementary semantic signal rather than a primary decision-making pathway.

Evaluation Metrics. Following the official shared-task setting, we use Macro-F1 as the primary evaluation metric. Macro-F1 computes the F1 score for each class separately and then averages them across classes, thereby assigning equal

Table 3: Hyperparameters used across the main backbone variants.

Parameter	Value
Backbone Variants	CLIP ViT-L/14; EVA02-L-14
Max Token Length	77
Batch Size	16
Max Epochs	10
Optimizer	AdamW
Learning Rate	3×10^{-5}
Weight Decay	5×10^{-4}
Label Smoothing	0.05
Residual Mixing Ratio	0.2
Cosine Classifier Scale	30

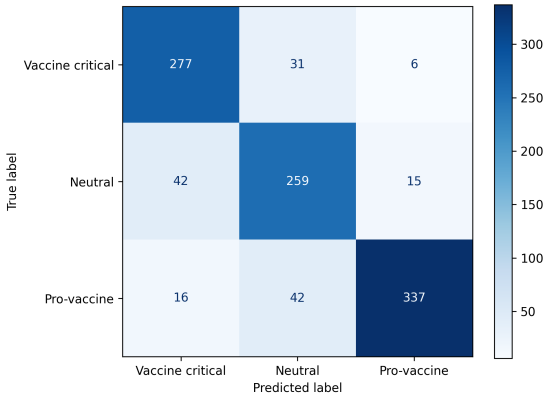

Figure 3: Confusion matrix of the final ensemble system on the official test set.

importance to *Vaccine critical*, *Neutral*, and *Pro-vaccine*. This metric is particularly suitable for the current task because it better reflects balanced classification performance across different stance categories and is less likely to be dominated by relatively easier or more frequent classes.In addition to Macro-F1, we also report Accuracy, Precision, and Recall to provide a more comprehensive view of system behavior. Accuracy reflects the overall proportion of correctly classified samples, while Precision and Recall help us analyze whether the model tends to be overly conservative or overly aggressive for certain classes. Reporting these auxiliary metrics allows us to better understand the trade-offs among overall correctness, class-wise reliability, and class-wise coverage, and also provides additional evidence for interpreting error patterns and comparing model variants.

6 Results and Discussion

6.1 Experimental Results

Our final submission achieved a Macro-F1 score of 0.8494 on the official test set, securing the first place in the shared task (Thapa et al., 2026b). Table 2 summarizes the empirical results of our primary models and their respective variants. The experimental findings indicate that the task-adapted MemeCLIP establishes a robust baseline on the VaxMeme dataset. Building upon this foundation, the integration of a more advanced backbone, noise-aware training, retrieval augmentation, and multi-model ensembling yields substantial performance gains. Notably, noise-aware weighted training contributes significantly to these improvements, suggesting that inherently noisy or highly ambiguous instances profoundly affect training stability within this specific context.

Conversely, the variant incorporating EVA-CLIP+Qwen auxiliary descriptions registers a noticeable performance degradation compared to alternative strategies. This observation implies that while LLM-generated descriptions theoretically enrich the semantic context for samples with poor OCR quality, these supplementary features do not consistently translate into classification improvements under the current experimental setup. This discrepancy may be attributed to residual noise within the generated text, semantic misalignment between the auxiliary descriptions and the original image-text pairs, or the model's suboptimal utilization of the supplementary semantic branch. Ultimately, the superior performance of the ensemble framework over all standalone configurations underscores a strong complementarity among the diverse variants. Techniques such as noise-aware training and text-source disentanglement prove to be non-redundant; rather, they synergistically provide comprehensive and robust evidence for decision-making during the ensembling phase.

6.2 Further Analysis

To further elucidate the limitations of our system, we conduct a detailed error analysis based on the confusion matrix. The most prominent failure mode involves the misclassification between the Neutral category and stance-bearing classes. This phenomenon suggests that the model is prone to over-inferring stances when processing memes characterized by purely factual content, weak opinion signals, or insufficient background context. Furthermore, instances degraded by low-quality OCR extraction, severe textual noise, or intricate visual layouts exhibit a significantly higher susceptibility to misclassification. Another intrinsic challenge arises from samples with inherently ambiguous

class boundaries, such as memes that ostensibly share objective information but implicitly convey supportive or critical nuances within specific socio-cultural contexts. Collectively, these error patterns reveal that while the proposed system effectively integrates multimodal signals, it retains pronounced limitations in handling weak-evidence scenarios, modeling implicit pragmatics, and resolving fine-grained semantic boundaries.

7 Conclusion

In this paper, we describe our proposed approach for the Shared Task on Multimodal Vaccine Critical Meme Detection (VaxMeme) @EEUCA 2026. Using MemeCLIP as the foundational framework, we systematically enhance the model through task-specific adaptation, noise-aware training, auxiliary semantic injection, text-source disentanglement, and inference-stage optimization. Our final system, constructed by ensembling multiple complementary model variants, achieves highly competitive performance on the official shared-task test set. Experimental results demonstrate the effectiveness of noise-aware training and late-stage ensembling strategies for robust vaccine meme stance identification. Future work will focus on developing more robust architectures and exploring interpretable multimodal reasoning mechanisms for complex meme analysis.

8 Limitations

Despite integrating multiple complementary variants, the system inherently operates within a supervised classification paradigm and lacks explicit modeling of nuanced socio-cultural contexts. Furthermore, while the LLM-generated auxiliary descriptions are designed for neutral and selective utilization, they may occasionally introduce semantic noise or over-interpretation. The retrieval-augmented module is also constrained by its reliance on semantically analogous instances in the training corpus, rendering it less robust when processing rare or out-of-distribution memes. Moreover, the current findings are primarily validated on the VaxMeme benchmark, and their generalizability to alternative public health domains, diverse social media platforms, or cross-cultural scenarios requires further empirical verification. Achieving optimal performance necessitates model ensembling, a strategy that inevitably introduces higher inference complexity and deployment costs.

References

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, and 8 others. 2025. Qwen2.5-VL technical report. *Preprint*, arXiv:2502.13923.

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*, pages 1994–2003.

Mingxuan Chen, Xinqiao Chu, and K. P. Subbalakshmi. 2021. MMCoVaR: Multimodal COVID-19 vaccine focused data repository for fake news detection and a baseline architecture for classification. In *Proceedings of the 2021 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, pages 31–38.

Zhi-Hao Guan, Qing-Yuan Jiang, and Yang Yang. 2025. Balance-aware sequence sampling makes multi-modal learning better. In *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence*, pages 2842–2850.

Ali Hürriyetoğlu, Surendrabikram Thapa, Hristo Tanev, Laxmi Thapa, and Surabhi Adhikari. 2026. Overview of the workshop on event extraction and understanding: Challenges and applications. In *Proceedings of the 9th Workshop on Event Extraction and Understanding: Challenges and Applications (EEUCA)*.

Qing-Yuan Jiang, Zhouyang Chi, and Yang Yang. 2025. Interactive multimodal learning via flat gradient modification. In *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence*, pages 5489–5497. ijcai.org.

Bin Liang, Ang Li, Jingqian Zhao, Lin Gui, Min Yang, Yue Yu, Kam-Fai Wong, and Ruifeng Xu. 2024. Multi-modal stance detection: New datasets and model. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 12373–12387, Bangkok, Thailand. Association for Computational Linguistics.

Usman Naseem, Jinman Kim, Matloob Khushi, and Adam G. Dunn. 2023. A multimodal framework for the identification of vaccine critical memes on twitter. In *Proceedings of the Sixteenth ACM International Conference on Web Search and Data Mining*, pages 706–714.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, pages 8748–8763.

Siddhant Bikram Shah, Shuvam Shiwakoti, Maheep Chaudhary, and Haohan Wang. 2024. MemeCLIP: Leveraging CLIP representations for multimodal meme classification. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 17320–17332, Miami, Florida, USA. Association for Computational Linguistics.

Quan Sun, Yuxin Fang, Ledell Wu, Xinlong Wang, and Yue Cao. 2023. EVA-CLIP: Improved training techniques for CLIP at scale. *arXiv preprint arXiv:2303.15389*.

Laxmi Thapa, Aryaman Jain, Lakshmojee Koduru, Surabhi Adhikari, Junaid Rashid, Jungeun Kim, Surendrabikram Thapa, and Usman Naseem. 2026a. Concept-grounded detection of vaccine misinformation in multimodal content using interpretable vision-language models. In *Companion Proceedings of the ACM Web Conference 2026*.

Surendrabikram Thapa, Siddhant Bikram Shah, Kritesh Rauniyar, Shuvam Shiwakoti, Surabhi Adhikari, Hariram Veeramani, Kristina T. Johnson, Ali Hurriyetoglu, Hristo Tanev, and Usman Naseem. 2025. Multimodal hate, humor, and stance event detection in marginalized sociopolitical movements. In *Proceedings of the 8th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Texts*, pages 20–31, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

Surendrabikram Thapa, Shuvam Shiwakoti, Siddhant Bikram Shah, Kritesh Rauniyar, Laxmi Thapa, Surabhi Adhikari, Kristina T. Johnson, Ali Hürriyetoğlu, Hristo Tanev, and Usman Naseem. 2026b. Multimodal identification of vaccine content stance on social media. In *Proceedings of the 9th Workshop on Event Extraction and Understanding: Challenges and Applications (EEUCA)*.

Yang Yang, Fengqiang Wan, Qing-Yuan Jiang, and Yi Xu. 2024. Facilitating multimodal classification via dynamically learning modality gap. In *Advances in Neural Information Processing Systems*, volume 37, pages 62108–62122. Curran Associates, Inc.

Feng Yu, Xiangyu Wu, Yang Yang, and Jianfeng Lu. 2026. Multimodal classification via total correlation maximization. In *The Fourteenth International Conference on Learning Representations*.