Hateful Meme Challenge with Tiny Reasoning

Sungjin Park
Sungkyunkwan University
sungjin.code@gmail.com

Abstract

Hateful meme detection is a challenging multimodal task, as harmful intent often arises from subtle interactions between images and text rather than from either modality alone. This requires contextual reasoning to distinguish benign confounders, where small modality-level changes can drastically alter semantic meaning. To address the limitations of feed-forward multimodal models that rely on shallow correlations, we propose a Multimodal Recursive Reasoning Framework for hateful meme detection. The framework builds on a pretrained CLIP encoder for strong joint representations and introduces a parameter-efficient Tiny Recursive Model that iteratively refines latent representations through recursive inference. We further adopt a hierarchical Recursive Learning Strategy, combining latent recursion, deep recursion, and deep supervision, to achieve effective reasoning depth without excessive memory overhead. Experiments on the Hateful Memes Challenge dataset demonstrate that recursive reasoning improves prediction balance and achieves the highest F1-score among compared methods, highlighting the effectiveness of parameter-efficient recursive inference for nuanced multimodal understanding.

1. Introduction

Detecting hateful content in internet memes is particularly challenging due to their inherently multimodal and context-dependent nature. Unlike standalone text or images, memes convey meaning through interactions between visual content and overlaid text, where neither modality alone is sufficient to determine harmful intent. Minimal changes in either modality can invert semantic interpretation, producing benign confounders [3] (Figure 1). Consequently, models that rely on unimodal cues or shallow correlations often fail to generalize, making hateful meme detection substantially more difficult than conventional hate speech classification. These characteristics suggest that the task fundamentally requires reasoning over joint visual–linguistic context rather than simple pattern matching.

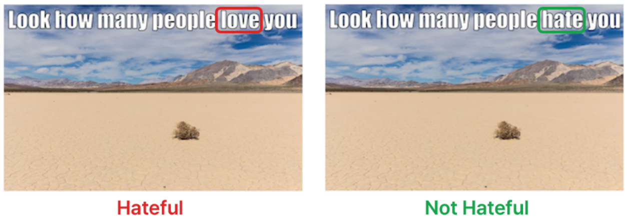


Figure 1. Example of confounder which is a minimal replacement of either the image or text that flips the label.

Correct classification frequently depends on understanding implicit assumptions, social stereotypes, or contextual contradictions that emerge only when visual and linguistic information are considered together [9]. Even strong multimodal encoders paired with feed-forward classifiers tend to collapse rich contextual signals into a single inference step, resulting in brittle decision boundaries [6]. This limitation motivates the need for architectures that can iteratively refine internal representations before producing a final prediction.

Recent work in multimodal learning increasingly explores reasoning-oriented architectures, where latent-space recurrence and parameter-efficient recursion enable deeper inference without prohibitive computational cost [12, 2]. These approaches suggest that structured recursive reasoning is well suited for tasks requiring nuanced semantic alignment under limited supervision.

Motivated by these observations, this paper proposes a Multimodal Recursive Reasoning Framework for hateful meme detection. The framework combines a pretrained CLIP encoder for robust multimodal feature extraction with a Tiny Recursive Model (TRM) that performs iterative reasoning in latent space. To support deep yet memory-efficient inference, we introduce a hierarchical Recursive Learning Strategy incorporating latent recursion, deep recursion with stop-gradient control, and deep supervision. This design enables substantial effective depth without the memory overhead of full backpropagation through long reasoning chains.

We evaluate the proposed approach on the Hateful

Memes Challenge dataset, which is explicitly constructed to discourage unimodal shortcuts through benign confounders. Experimental results show that while CLIP-based models already provide strong baselines, incorporating recursive reasoning consistently improves prediction balance, as reflected by higher F1-scores. These findings demonstrate that parameter-efficient recursive inference plays a critical role in robust multimodal hateful meme detection and highlight the importance of structured reasoning mechanisms for reliable content moderation.

2. Related work

2.1. Hateful meme detection

Recent studies on meme analysis emphasizes that understanding the interaction between the two modalities is essential and has proposed a variety of benchmark datasets accordingly. Suryawanshi et al. [10] constructed the MultiOFF dataset using memes related to the 2016 U.S. presidential election to detect offensive content. Their study demonstrated that early fusion models combining textual and visual information outperform unimodal approaches. Nevertheless, existing CLIP-based models exhibit limitations in distinguishing so-called confounder memes, in which subtle differences between image and text lead to entirely different semantics. To address this issue, Mei et al. [6] proposed Retrieval-Guided Contrastive Learning. This approach dynamically retrieves semantically similar positive examples and hard negative examples with opposite hate labels during training, enabling contrastive learning that encourages the model to capture fine-grained features associated with hateful content. Experimental results show that this method outperforms large-scale multimodal models on the HatefulMemes dataset.

As the number of parameters in pre-trained vision-language models (PVLM) increases, methodologies that efficiently exploit pretrained knowledge without full model fine-tuning have gained attention. Cao et al. [1] investigated an approach that converts meme images into textual captions and feeds them as prompts to language models. Rather than relying on generic caption generation, they proposed the Pro-Cap method, which employs a visual question answering paradigm to generate captions by asking questions about critical attributes for hate detection, such as race and gender. This approach achieves strong detection performance using frozen PVLMs. Recent studies further extend beyond simple binary classification toward an Explain-then-Detect paradigm, in which models are required to justify their decisions. Inspired by how human reviewers learn policy manuals, Mei et al. [7] proposed ExPO-HM, which combines policy-based supervised fine-tuning warm-up with curriculum-based reinforcement learning. To enhance the quality of explanations, conditional decision entropy is incorporated as a reward function. In addition, to enable detection in zero-shot settings without labeled training data, Liu et al. [5] introduced MIND, a multi-agent framework. MIND retrieves similar memes from an unlabeled reference dataset, derives insights from them, and conducts inter-agent debates to reach a final judgment on harmfulness. This framework provides robust detection performance without continuous data labeling, aligning well with the rapidly evolving nature of internet memes.

2.2. Reasoning model

Efforts to overcome the reasoning limitations of large language models (LLMs) by emulating the operational principles of the human brain have recently been concretized through the proposal of the Hierarchical Reasoning Model. The HRM [12] is inspired by the brain's multi-timescale processing and hierarchical organization, and adopts a dual recurrent neural network architecture that combines a high-level module responsible for slow and abstract planning with a low-level module that performs fast and concrete computations. To address the data inefficiency inherent in conventional Chain-of-Thought approaches, HRM conducts iterative reasoning within a latent space and supervises intermediate states through deep supervision. In particular, HRM introduces a one-step gradient approximation and Adaptive Computational Time, thereby mitigating the memory overhead of Backpropagation Through Time and demonstrating superior performance over existing LLMs on complex reasoning tasks such as ARC-AGI and Sudoku-Extreme, using only 27 million parameters and 1,000 training examples.

Building on these results, Jolicoeur-Martineau [2] recently advanced the efficiency of recursive reasoning by proposing the Tiny Recursive Model, which addresses the complex biological assumptions and structural inefficiencies of HRM. TRM critiques HRM's reliance on gradient approximation via fixed-point theory and its use of two separate networks, showing instead that stronger generalization can be achieved with a single tiny network. It simplifies HRM's hierarchical latent variables into a more intuitive formulation consisting of the current output and a latent reasoning state, and replaces the costly Adaptive Computational Time mechanism with a single forward pass to improve training efficiency. As a result, TRM achieves over 30% higher accuracy than HRM on Sudoku-Extreme and improved performance on ARC-AGI using only two layers and seven million parameters, thereby highlighting new possibilities for parameter-efficient recursive reasoning.

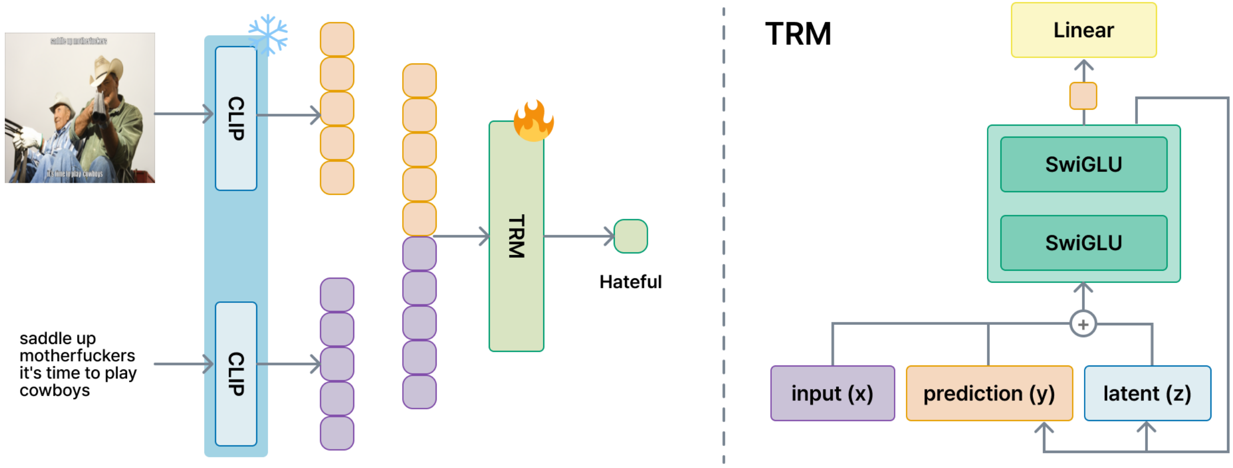


Figure 2. Overview of the proposed architecture. The model comprises a CLIP for extracting visual and linguistic features, and a TRM that conducts deep reasoning over the representations.

3. Methods

3.1. Overall Architecture

This study proposes a Multimodal Recursive Reasoning Framework designed to capture complex interactions between images and text and to perform advanced binary classification based on such interactions. The proposed model consists of a multimodal feature encoder, which extracts visual and linguistic features, and a Tiny Recursive Model that conducts deep reasoning over the extracted representations as illustrated in Figure 2. To effectively embed visual and textual information from the input data, a pretrained CLIP (Contrastive Language–Image Pretraining) model is employed as the backbone. The input image I and text T are independently processed by the CLIP encoders and transformed into high-dimensional feature vectors. These modality-specific feature vectors are then concatenated to form a unified input representation $\mathbf{x}$.

To overcome the reasoning limitations of a simple feed-forward architecture, this work introduces the TRM architecture to perform iterative refinement within a latent space. The TRM takes as input the static representation $\mathbf{x}$ together with two dynamically updated state vectors at each step: the prediction state $\mathbf{y}_t$ and the latent reasoning state $\mathbf{z}_t$. At reasoning step t, computation is performed by first fusing the three vectors through element-wise summation, followed by a nonlinear transformation. To maximize representational capacity while maintaining parameter efficiency, an MLP layer incorporating the SwiGLU (Swish-Gated Linear Unit) activation function is used:

$$\mathbf{h}_t = \mathrm{SwiGLU}(\mathbf{x} + \mathbf{y}_{t-1} + \mathbf{z}_{t-1})$$

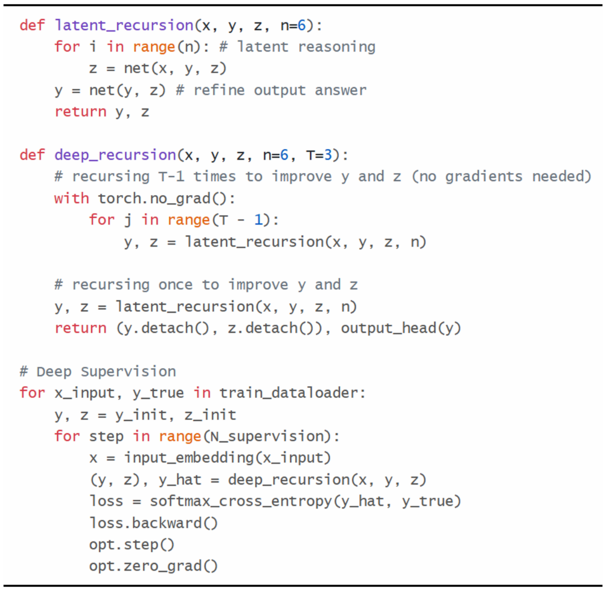
Here, $\mathbf{z}$ serves to deepen the reasoning process by refining latent information based on the input context $\mathbf{x}$ and the current latent state, while $\mathbf{y}$ concretizes the final prediction using the refined latent representation $\mathbf{z}$. This procedure is repeated for N recursive iterations, enabling the model to achieve substantial effective depth despite a limited number of parameters. After the recursive reasoning process is completed, the final prediction state vector $\mathbf{y}_N$ is passed to a linear classifier. This linear layer maps the condensed multimodal representation to logits corresponding to the two target classes, from which the final prediction probabilities are obtained.

3.2. Reasoning Process

We adopt a Recursive Learning Strategy that progressively deepens reasoning by leveraging the static context vector $\mathbf{x}$ extracted from the CLIP encoder together with two dynamic states: the prediction vector $\mathbf{y}$ and the latent vector $\mathbf{z}$. The overall learning process follows a hierarchical structure consisting of (1) Latent Recursion, (2) Deep Recursion, and (3) Deep Supervision. Figure 3 illustrates the pseudocode of our method.

3.2.1 Latent Recursion

The latent recursion process, which constitutes the basic unit of reasoning, decouples exploration in the latent space from prediction refinement. Specifically, given the context $\mathbf{x}$ and the current prediction $\mathbf{y}$ held fixed, the model repeatedly updates the latent state $\mathbf{z}$ for n iterations. This process, referred to as latent reasoning, encourages the model to sufficiently internalize the problem context through itera-



```
def latent_recursion(x, y, z, n=6):
    for i in range(n): # latent reasoning
        z = net(x, y, z)
    y = net(y, z) # refine output answer
    return y, z

def deep_recursion(x, y, z, n=6, T=3):
    # recursing T-1 times to improve y and z (no gradients needed)
    with torch.no_grad():
        for j in range(T - 1):
            y, z = latent_recursion(x, y, z, n)

    # recursing once to improve y and z
    y, z = latent_recursion(x, y, z, n)
    return (y.detach(), z.detach()), output_head(y)

# Deep Supervision
for x_input, y_true in train_dataloader:
    y, z = y_init, z_init
    for step in range(N_supervision):
        x = input_embedding(x_input)
        (y, z), y_hat = deep_recursion(x, y, z)
        loss = softmax_cross_entropy(y_hat, y_true)
        loss.backward()
        opt.step()
        opt.zero_grad()
```

Figure 3. Pseudocode of the proposed Recursive Learning Strategy. The algorithm illustrates the hierarchical learning process comprising Latent Recursion, Deep Recursion, and Deep Supervision. It details how the static context $\mathbf{x}$ and dynamic states $\mathbf{y}, \mathbf{z}$ are utilized to progressively deepen reasoning and refine predictions.

tive internal computation before committing to a final decision. After completing n latent reasoning steps, the model performs a single prediction refinement step, updating the prediction state $\mathbf{y}$ based on the refined latent representation $\mathbf{z}$.

$$\begin{aligned}\mathbf{z}^{(i)} &= \mathcal{F}(\mathbf{z}^{(i-1)}, \mathbf{x} + \mathbf{y}) \quad \text{for } i = 1 \dots n \\ \mathbf{y}' &= \mathcal{F}(\mathbf{y}, \mathbf{z}^{(n)})\end{aligned}$$

Here, $\mathcal{F}$ denotes the TRM network incorporating a SwiGLU block.

3.2.2 Deep Recursion

To increase reasoning depth while maintaining memory efficiency during training, a deep recursion strategy is introduced. In this setting, the latent recursion block is executed a total of T times. As a key training mechanism, the first $T - 1$ iterations are performed with gradient computation disabled, resulting in inference-only forward passes. Gradients are computed only in the final T-th iteration, allowing the model to preserve contextual information across a deep computational graph while substantially reducing backpropagation overhead.

The final training procedure follows a Deep Supervision scheme over N_{sup} stages. At each supervision step, the input $\mathbf{x}$ remains unchanged, while the prediction and latent states $\mathbf{y}$ and $\mathbf{z}$ produced in the previous step are reused as initial states. A strategy analogous to Truncated Backpropagation Through Time is applied, such that $\mathbf{y}$ and $\mathbf{z}$ are detached from the computational graph at the end of each step. By immediately computing the loss and performing backpropagation for each predicted output $\hat{\mathbf{y}}_k$, the model receives stable feedback throughout the entire reasoning process and progressively converges toward the correct solution.

To enable efficient learning of the reasoning process, we employ a Deep Supervision strategy combined with iterative parameter updates. Instead of accumulating gradients over the entire recursion, we compute the loss and perform an optimization step at each supervision step k. The step-wise loss function $\mathcal{L}_k$ is defined as:

$$\mathcal{L}_k = -\left[y_{true} \cdot \log(\sigma(\hat{y}_k)) + (1 - y_{true}) \cdot \log(1 - \sigma(\hat{y}_k))\right]$$

where $y_{true} \in \{0, 1\}$ denotes the ground-truth label, $\hat{y}_k$ is the logit predicted at step k, and $\sigma(\cdot)$ represents the sigmoid function. At each step k, the model parameters are updated to minimize $\mathcal{L}_k$, and the recursive states y_k and z_k are detached from the computational graph. This approach effectively implements Truncated Backpropagation Through Time, allowing the model to refine its predictions sequentially while maintaining training stability.

4. Experiments

4.1. Dataset

This study evaluates the proposed model's multimodal hate speech detection performance using the Hateful Memes Challenge Dataset[3]. This dataset is a benchmark constructed to identify hateful expressions embedded in multimodal memes on social media, where correct classification requires understanding the contextual meaning that emerges from the combination of text and image. The full dataset consists of 10,000 memes. However, because the labels of the official test set released for the 2020 challenge remain private, the official validation set of 500 samples is used as a substitute test set for performance evaluation in this study. For model training and validation, the original training set of 8,500 samples is randomly re-split with an 8:2 ratio, resulting in 6,800 training samples, 1,700 validation samples, and 500 test samples used in the experiments.

A defining characteristic of this dataset is the inclusion of benign confounders, which are designed to prevent models from solving the task by relying solely on unimodal information. Benign confounders are constructed by minimally altering either the image or the text in a hateful meme to flip its label to non-hateful. For instance, a given text may be considered hateful when paired with a specific image, but non-hateful when combined with a neutral image.

Such a design forces models to move beyond superficial biases—such as the presence of particular words or visual objects—and instead perform subtle reasoning based on the interaction between image and text modalities.

4.2. Implementation Details

We implemented our framework using PyTorch on Ubuntu 24.04 LTS, equipped with a single NVIDIA RTX 4060 GPU and 32GB of RAM. For optimization, we utilized the AdamW optimizer with a batch size of 64. The initial learning rate was set to 1×10^{-4} with a weight decay of 5×10^{-2}. We employed a cosine annealing scheduler with a linear warmup phase. The model was trained for a maximum of 30 epochs, incorporating an early stopping mechanism to prevent overfitting. Data augmentation was performed using TrivialAugmentWide to enhance model generalization. Regarding the specific hyperparameters of our proposed architecture, the number of supervision signals (N_{sup}) was set to 8, and the recursion parameters ($recursion\_n$ and $recursion\_t$) were both set to 3.

4.3. Evaluation Metrics

This study employs three primary metrics to provide a comprehensive analysis of model performance. First, AUC (Area Under the Curve), which serves as the main evaluation metric and the ranking criterion of the challenge, is adopted. AUC reflects the stability of model performance across varying decision thresholds and is well suited for assessing discriminative capability in binary classification tasks such as hate speech detection. Second, accuracy is reported to facilitate intuitive interpretation of performance. Accuracy represents the proportion of correctly classified samples over the entire test set and provides an overall view of predictive capability. Third, the F1-score is included as an additional metric. As the harmonic mean of precision and recall, the F1-score evaluates the balance between predictions for hateful and non-hateful classes that may not be fully captured by accuracy alone. This metric is particularly useful for verifying whether the model produces robust and unbiased predictions across classes.

4.4. Comparison with Previous Methods

We selected the baseline model released by Facebook AI and several top-performing models from the challenge leaderboard, as summarized in Table 1. As discussed earlier, because the labels of the official test set are not publicly available, our comparison is conducted under identical conditions by directly referencing the performance reported on the original validation set in prior studies. The reported values are adopted without modification. Consequently, some evaluation metrics are left unreported for certain methods, reflecting the fact that those scores were not provided in the corresponding prior work. The results indicate that the proposed models based on a pretrained CLIP encoder consistently outperform existing baselines.

Table 1. Comparison with previous models on the dataset.

Models	**AUC**	**Accuracy**	**F1**
ViLBERT CC [3]	70.8	70.4	
Visual BERT COCO [3]	73.7	70.8	
VILIO [8]	81.6	-	
Visual BERT [11]	75.2	71.0	
UNITER [4]	79.1	-	
CLIP + MLP (Ours)	**82.6**	**75.4**	65.8
CLIP + TRM (Ours)	81.9	72.7	**67.6**

First, the CLIP + MLP model achieves the highest performance among the compared methods, recording an AUC of 82.6 and an accuracy of 75.4%. This result demonstrates that large-scale multimodal pretraining with CLIP provides strong feature representations for tasks such as meme analysis, where visual and linguistic information are tightly coupled. This relatively simple architecture exhibits superior generalization performance.

Second, the CLIP + TRM model achieves an AUC of 81.9 and an accuracy of 72.7%, showing a slight decrease relative to CLIP + MLP. However, it achieves the highest F1-score of 67.6, compared to 65.8 for CLIP + MLP, corresponding to an improvement of approximately 1.8%p. In tasks such as the Hateful Memes dataset, where label imbalance and nuanced semantic interpretation are critical, improvements in F1-score are particularly meaningful. These results suggest that the recursive reasoning process introduced by TRM enables the model to move beyond majority-class prediction and to more precisely explore the decision boundary between hateful and non-hateful classes, effectively balancing precision and recall.

Overall, the strong representational capacity of CLIP substantially enhances detection performance, and the integration of the TRM module further enables more reliable and balanced hate speech detection by mitigating class bias and refining multimodal reasoning.

5. Discussion and Future Work

Although the proposed Multimodal Recursive Reasoning Framework demonstrates competitive performance, the experimental results indicate that the integration of the TRM does not consistently outperform simpler feed-forward alternatives in terms of overall AUC and accuracy. A key factor underlying this limitation appears to be the insufficient contextual richness of the input representations provided to the recursive reasoning module. Specifically, TRM is designed to iteratively refine predictions based on a fixed static context $\mathbf{x}$ extracted by the multimodal feature encoder. Although CLIP provides strong visual–linguistic alignment, its representations are primarily grounded in perceptual and lexical correlations learned during pretraining. As a result, certain memes that rely on high-context knowledge—such as contemporary social issues, historical references, political events, or culturally specific stereotypes—may not be adequately encoded at the feature extraction stage. In such cases, recursive reasoning operates in an incomplete or impoverished context, limiting its ability to fully exploit the advantages of iterative inference. This observation suggests that recursive reasoning alone is not sufficient; its effectiveness is fundamentally constrained by the semantic completeness of the underlying representations. Nevertheless, the improved F1-score observed with TRM indicates that recursive reasoning contributes to more balanced decision-making, particularly in ambiguous cases near the class boundary. This implies that TRM is capable of refining subtle distinctions when relevant information is present, but struggles when critical external context is absent from the input space.

These findings point to several promising directions for future work. First, incorporating external knowledge sources—such as retrieval-augmented textual context, structured knowledge bases, or up-to-date web-derived information—could enrich the static context $\mathbf{x}$ and allow reasoning modules to operate over more semantically complete representations. Designing mechanisms that allow such high-context information to be dynamically injected into the reasoning loop would be particularly beneficial for memes that depend on evolving social or political discourse. Second, future research may explore knowledge-aware feature encoders or hybrid architectures that jointly learn perceptual grounding and abstract world knowledge. Finally, extending the recursive framework toward explanation-aware or agent-based reasoning may further enhance interpretability and robustness in real-world content moderation scenarios.

Overall, this discussion highlights that effective multimodal reasoning is a joint function of both inference architecture and contextual coverage, and that enriching the latter is essential to fully realize the potential of recursive reasoning models.

6. Conclusion

This work introduced a Multimodal Recursive Reasoning Framework for hateful meme detection that explicitly models cross-modal interactions through iterative latent reasoning. By combining a pretrained CLIP encoder with a parameter-efficient Tiny Recursive Model, the proposed approach overcomes the limitations of conventional feed-forward architectures that rely on shallow multimodal correlations.

A hierarchical Recursive Learning Strategy enables deep yet memory-efficient reasoning, ensuring stable training while achieving substantial effective depth. Experimental results on the Hateful Memes Challenge dataset show that incorporating recursive reasoning improves prediction balance and yields the highest F1-score among compared methods. While simpler CLIP-based models achieve strong accuracy, recursive inference provides more reliable decision boundaries in the presence of multimodal confounders.

Overall, these findings highlight parameter-efficient recursive reasoning as an effective and scalable solution for multimodal hate speech detection, with promising potential for broader multimodal reasoning tasks that require nuanced semantic understanding.

References

[1] R. Cao, M. S. Hee, A. Kuek, W.-H. Chong, R. K.-W. Lee, and J. Jiang. Pro-cap: Leveraging a frozen vision-language model for hateful meme detection. In *Proceedings of the 31st ACM International Conference on Multimedia*, MM '23, page 5244–5252, New York, NY, USA, 2023. Association for Computing Machinery.

[2] A. Jolicoeur-Martineau. Less is more: Recursive reasoning with tiny networks. *arXiv preprint arXiv:2510.04871*, 2025.

[3] D. Kiela, H. Firooz, A. Mohan, V. Goswami, A. Singh, P. Ringshia, and D. Testuggine. The hateful memes challenge: Detecting hate speech in multimodal memes. *Advances in neural information processing systems*, 33:2611–2624, 2020.

[4] P. Lippe, N. Holla, S. Chandra, S. Rajamanickam, G. Antoniou, E. Shutova, and H. Yannakoudakis. A multimodal framework for the detection of hateful memes. *arXiv preprint arXiv:2012.12871*, 2020.

[5] Z. Liu, C. Fan, H. Lou, Y. Wu, and K. Deng. MIND: A multi-agent framework for zero-shot harmful meme detection. In W. Che, J. Nabende, E. Shutova, and M. T. Pilehvar, editors, *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 923–947, Vienna, Austria, July 2025. Association for Computational Linguistics.

[6] J. Mei, J. Chen, W. Lin, B. Byrne, and M. Tomalin. Improving hateful meme detection through retrieval-guided contrastive learning. In L.-W. Ku, A. Martins, and V. Srikumar, editors, *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5333–5347, Bangkok, Thailand, Aug. 2024. Association for Computational Linguistics.

[7] J. Mei, M. Sun, J. Chen, P. Qin, Y. Li, D. Chen, and B. Byrne. Expo-hm: Learning to explain-then-detect for hateful meme detection. *arXiv preprint arXiv:2510.08630*, 2025.

[8] N. Muennighoff. Vilio: State-of-the-art visio-linguistic models applied to hateful memes. *arXiv preprint arXiv:2012.07788*, 2020.

[9] S. Pramanick, S. Sharma, D. Dimitrov, M. S. Akhtar, P. Nakov, and T. Chakraborty. MOMENTA: A multimodal framework for detecting harmful memes and their targets. In M.-F. Moens, X. Huang, L. Specia, and S. W.-t. Yih, editors, *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 4439–4455, Punta Cana, Dominican

Republic, Nov. 2021. Association for Computational Linguistics.

[10] S. Suryawanshi, B. R. Chakravarthi, M. Arcan, and P. Buitelaar. Multimodal meme dataset (MultiOFF) for identifying offensive content in image and text. In R. Kumar, A. K. Ojha, B. Lahiri, M. Zampieri, S. Malmasi, V. Murdock, and D. Kadar, editors, *Proceedings of the Second Workshop on Trolling, Aggression and Cyberbullying*, pages 32–41, Marseille, France, May 2020. European Language Resources Association (ELRA).

[11] R. Velioglu and J. Rose. Detecting hate speech in memes using multimodal deep learning approaches: Prize-winning solution to hateful memes challenge. *arXiv preprint arXiv:2012.12975*, 2020.

[12] G. Wang, J. Li, Y. Sun, X. Chen, C. Liu, Y. Wu, M. Lu, S. Song, and Y. A. Yadkori. Hierarchical reasoning model. *arXiv preprint arXiv:2506.21734*, 2025.

A. Declaration of Generative AI

During the preparation of this work, the authors used ChatGPT in order to translate the manuscript. After using this tool/service, the authors reviewed and edited the content as needed.