A Multimodal Hate Speech Detection Framework Based on Multi-level Cross-modal Attention and Gated Fusion

Rui Lv, Jie Wang, Xuan Liu, Lirong Chen†

College of Computer Science (College of Software), College of Artificial Intelligence, Inner Mongolia University, Hohhot 010000, Inner Mongolia, China

Abstract

Multimodal hate speech detection aims to integrate various modalities—such as text and images—to identify complex and implicit hateful content, thereby contributing to a healthier online environment. Despite notable progress in fusion techniques, existing approaches still struggle with modeling both local and global semantics and achieving effective cross-modal integration. To address these limitations, we propose MLCA, a novel multimodal hate speech detection framework. Our method employs a Twitter-based RoBERTa model and the Swin Transformer V2 to encode textual and visual modalities, respectively. These modality-specific representations are subsequently fused using a multi-level cross-modal attention mechanism. In addition, a dynamic gating module is introduced to adaptively integrate attention features across different semantic levels. We conduct comprehensive evaluations on two benchmark datasets and compare our model with a wide range of state-of-the-art unimodal and multimodal baselines. Experimental results show that our framework consistently surpasses state-of-the-art methods on both datasets.

Keywords: Multimodal hate speech detection; cross-modal attention; gated fusion; deep learning; multimodal fusion

1. Introduction

With the rapid proliferation of social media platforms, the spread of hate speech has emerged as an increasingly urgent societal concern[1]. Platforms such as Twitter are frequently exploited to disseminate hateful content[2]. Hate speech refers to discourse that targets individuals or groups based on race, ethnicity, gender, or religion[3], posing serious threats to both social cohesion and public safety[4].

Modern hate speech has evolved beyond plain text, with multimodal content—particularly the combination of text and images—becoming increasingly prevalent. Images not only enhance the expressiveness of textual messages but also serve as covert channels for conveying hateful intent[5]. The rise of visual-linguistic memes has amplified both the semantic complexity and subtlety of hate speech, thereby making it increasingly difficult to detect[6, 7].

To address these challenges, multimodal hate speech detection has emerged as a promising solution that integrates textual and visual signals to capture richer and more nuanced semantic

†Corresponding author: Lirong Chen (Email: lrchen10@126.com; ORCID: 0000-0003-4516-209X)

cues[8, 9]. Compared to unimodal methods, multimodal approaches offer enhanced representational capacity and greater robustness to subtle and implicit forms of hate. Recent research has primarily focused on improving feature fusion, which fundamentally relies on the accurate alignment of multimodal features across multiple semantic levels. Fusion strategies are generally categorized into early fusion, which integrates features during the encoding phase, and late fusion, which aggregates the outputs after each modality has been processed independently.

However, existing fusion strategies—whether early or late—remain insufficient for modeling the complex interactions between text and images that are essential for effective hate speech detection. Many methods rely on multilayer perceptrons or shallow attention mechanisms, which often fail to capture both fine-grained local features and global semantic dependencies. Furthermore, current frameworks struggle to dynamically assess the importance of semantic features across multiple levels, thereby diminishing the effectiveness of cross-modal coordination.

To overcome these limitations, we propose MLCA, a novel framework that combines multi-level cross-modal attention with a gated fusion strategy. MLCA extracts hierarchical features from text and images using RoBERTa and Swin Transformer, respectively, and facilitates deep semantic interaction within a shared latent space via multi-level cross-modal attention. A residual normalization mechanism is introduced to stabilize training and improve information flow. Finally, a gated fusion module adaptively integrates attention outputs from different levels according to their semantic contributions. The main contributions of our work are summarized as follows:

- We propose a multi-level cross-modal attention mechanism that progressively aligns global textual semantics with multi-scale visual features, enabling finer text-image interaction and improving intermediate fusion quality.
- We design a gated fusion module that adaptively integrates multi-level interaction features using learnable weights, enhancing the model's ability to capture salient information from complex hate memes.
- We evaluate our model on two benchmark datasets and demonstrate, through comprehensive experiments, that it outperforms existing state-of-the-art baselines in multimodal hate speech detection.

2. Related work

Our work primarily builds upon two major lines of research: unimodal hate speech detection and multimodal hate speech detection.

2.1. Unimodal Hate Speech Detection

Early research on hate speech detection primarily focused on explicit content, typically marked by overtly offensive or abusive language. Most studies adopted conventional classification pipelines leveraging BERT-family pre-trained models for sentence-level encoding. For example, HateXplain[10] provided fine-grained annotations and explainable labels for supervised learning. Masued et al.[11] further emphasized identifying explicit hate spans to improve model interpretability.

However, implicit hate speech presents greater challenges due to its subtle and indirect nature, often expressed through metaphor, sarcasm, or irony[12, 13]. Traditional models tend to underperform on such content, as demonstrated by the Implicit Hate Corpus[14]. To address this limitation, contrastive learning has been widely explored. ImpCon[15] constructed semantically similar pairs to guide models in distinguishing metaphorical expressions. SimCSE[16] leveraged natural language inference labels to generate sentence-level contrastive signals. Building upon these efforts, Lu et al.[17] proposed Dual Contrastive Learning, which aligns contrastive objectives among raw texts, pseudo-labels, and label semantics, significantly improving the recognition of metaphorical and sarcastic hate expressions.

2.2. Multimodal Hate Speech Detection

In recent years, hate speech on social media has increasingly adopted multimodal forms, evolving beyond text-only expressions to combinations of text and images—often embedded in memes or visual metaphors to obscure hateful intent[18]. This cross-modal complexity incorporates both explicit and implicit signals, thereby increasing the difficulty of accurate detection. As a result, multimodal hate speech detection has emerged as a critical research direction, aiming to integrate heterogeneous features for robust detection across diverse scenarios.

Among various modality combinations, image-text fusion has garnered the most attention. The Hateful Memes Challenge at NeurIPS 2020[19] established standardized benchmarks and evaluation metrics, significantly driving progress in the field. Its associated dataset, HatefulMemes, remains a widely used resource. Subsequently, Gomez et al.[20] introduced MMHS150K, a large-scale Twitter-derived dataset with rich inter-modal annotations that has since become widely used for evaluation.

Early multimodal approaches combined visual and textual features using traditional classifiers, such as logistic regression, thereby demonstrating the effectiveness of leveraging multimodal signals[21]. With the rise of deep learning, approaches have advanced to incorporate semantic fusion and cross-modal interaction mechanisms. For instance, Maity et al.[22] integrated sentiment and sarcasm cues for meme-level hate detection, while Lee et al.[23] proposed DisMultiHate, a disentangled framework that enhances interpretability via entity-level modeling.

Recent work further expands cross-modal reasoning capabilities. Cao et al.[24] utilized VQA-based image captioning to improve downstream understanding. Ayetiran and Özgöbek[25] introduced a unified model that integrates image, text, and embedded OCR features using cross-modal attention. Most notably, Xu et al. [26]proposed a prompt-based hypergraph fusion framework that enables structured reasoning over implicit cues and supports multi-target audience inference, achieving state-of-the-art performance on multiple benchmark datasets.

3. Methodology

3.1. Model Overview

The objective of multimodal hate speech detection is to identify diverse forms of hateful content conveyed through multiple modalities, such as text and images. Formally, a multimodal hate speech dataset $D = (X, Y)$ consists of paired samples (x_i, y_i), where $x_i \in X$ represents the multimodal input and $y_i \in Y$ denotes the corresponding ground-truth label. Each input x_i typically comprises a text component t_i and an image component m_i, forming a tuple $X = (T, I)$. The goal is to determine whether a given sample contains hateful content by jointly analyzing semantic cues from both modalities and predicting the corresponding hate label y.

To address the above challenges, we propose a Multimodal Hate Speech Detection Framework based on Multi-Level Cross-Modal Attention (MLCA). As illustrated in Figure 1, the framework consists of three key modules: a feature extraction module, a multi-level cross-modal attention fusion module, and a prediction module. The feature extraction module utilizes pre-trained RoBERTa and Swin Transformer V2 models to encode textual and visual inputs, respectively, generating global semantic embeddings and multi-scale visual features. The fusion module applies a hierarchical cross-modal attention mechanism to progressively perform feature fusion across multiple semantic levels of text and image representations. To improve the stability of this process, residual connections and layer normalization are incorporated as enhancement strategies. A gated fusion mechanism is then employed to dynamically aggregate the fused features, resulting in a unified multimodal representation. This final representation is subsequently passed into a multilayer perceptron classifier to perform hate speech prediction.

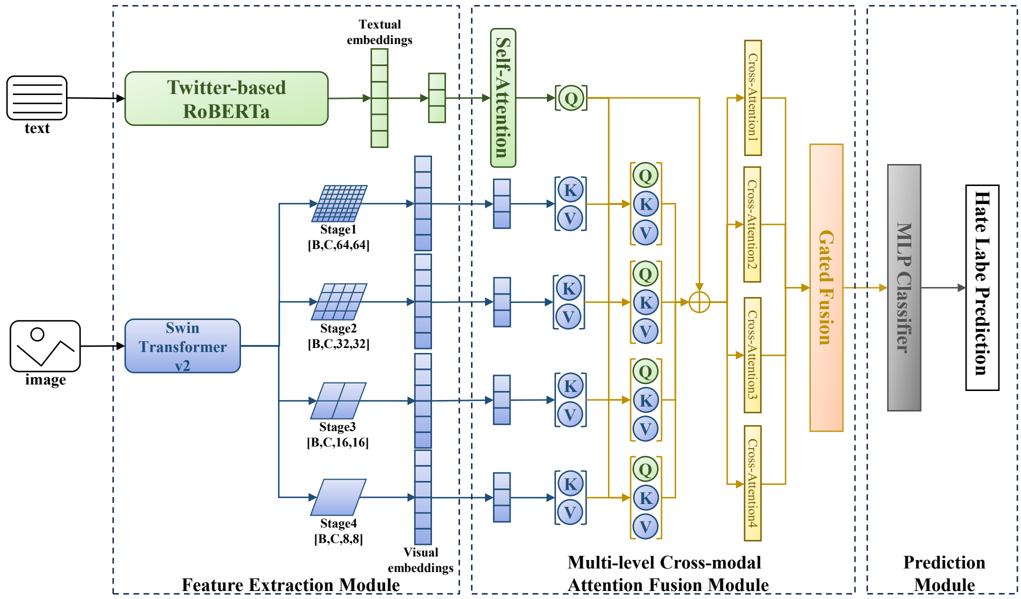

Figure 1: Overall Architecture of the Proposed MLCA Model

3.2. *Feature Extraction*

3.2.1. *Text Feature Extraction*

We adopt a Twitter-based RoBERTa model[27] as the text encoder. This model has been pre-trained on a corpus of 154 million tweets collected between January 2018 and December 2022, making it highly adaptable to the linguistic characteristics of social media. Given that texts in multimodal hate speech scenarios often exhibit properties such as short length, informal expressions, abbreviations, and slang (features commonly found in tweets), we employ this model to enhance semantic understanding of hateful intent. The Twitter-based RoBERTa model is publicly available via the Hugging Face Transformer API[1].

Given an input text t the model outputs the final hidden states H_t; we extract the [CLS] token representation and project it into a shared representation space to obtain the text feature vector H_t^{proj}.

[1]https://huggingface.co/cardiffnlp/twitter-roberta-large-2022-154m

3.2.2. Image Feature Extraction

We utilize Swin Transformer V2[28] as the image encoder. Specifically, we implement the swinv2_base_window16_256[2] variant via the TIMM framework, which is pre-trained on the ImageNet-1K dataset containing over 1.2 million labeled images spanning 1,000 categories. SwinV2 builds upon the original Swin Transformer[29] by introducing improved normalization strategies and enhanced model scaling, while maintaining its core window-based attention mechanism for efficient and scalable representation learning.

SwinV2 outputs four-stage feature maps corresponding to distinct semantic levels. Let the input image be m, and the flattened patch token sequence at each stage be denoted as F_i. We incorporate learnable positional encodings to preserve spatial information, as defined in Eq. 1:

$$F_i = \mathrm{Flatten}(SwinV2_i(m)) + P_i F_i = \mathrm{Flatten}(SwinV2_i(m)) + P_i \tag{1}$$

Each feature map F_i is then linearly projected into the shared representation space to obtain F_i^{proj}, which is used in subsequent cross-modal fusion.

3.3. Multi-Level Cross-Modal Attention Mechanism

To facilitate semantic alignment and interaction between textual and visual modalities at multiple levels, we design a multi-level cross-modal attention mechanism comprising three components: multi-scale cross-modal attention, residual normalization, and gated fusion. These components work in synergy to enable deep cross-modal interaction and dynamic integration.

3.3.1. Multi-Scale Cross-Modal Attention

To enhance the semantic representation of the text, we first apply multi-head self-attention over the projected text sequence H_t^{proj}, capturing intra-token dependencies. The output of the [CLS] token is used as the global semantic representation of the text, as shown in Eq. 2:

$$Q = \mathrm{MultiHeadSelfAttn}\left(H_t^{\mathrm{proj}}\right)[:, 0, :] \tag{2}$$

This vector Q serves as the query in the cross-modal attention mechanism, which interacts with image features F_i^{proj}, from each of the four SwinV2 stages. Cross-modal interaction is modeled via multi-head attention to yield fused representations, as defined in Eq. 3:

$$Z_i = \mathrm{MultiHeadAttn}\left(Q, F_i^{\mathrm{proj}}, F_i^{\mathrm{proj}}\right) \tag{3}$$

3.3.2. Residual Connection and Normalization Strategy

To stabilize deep cross-modal interactions and ensure smooth information flow, we apply residual connections and layer normalization to each attention output. Specifically, the cross-attention output Z_i is added to the query Q, regularized with DropPath, and normalized with LayerNorm, as shown in Eq. 4:

$$\tilde{Z}_i = \mathrm{LayerNorm}(\mathrm{DropPath}(Z_i) + Q) \tag{4}$$

Here, DropPath randomly drops connections during training to reduce overfitting, while LayerNorm ensures output stability and accelerates convergence.

[2] https://huggingface.co/timm/swinv2_base_window8_256.ms_in1k

3.3.3. Gated Fusion Strategy

To integrate information from different semantic levels, we design a gated fusion module that assigns learnable weights to the cross-modal outputs from each stage. The four intermediate outputs $\tilde{Z}_1$ to $\tilde{Z}_4$ are concatenated into a single vector:

$$Z_{concat} = [\tilde{Z}_1; \tilde{Z}_2; \tilde{Z}_3; \tilde{Z}_4] \quad (5)$$

This vector is passed through a linear transformation followed by softmax normalization to compute the attention weights α, as in Eq. 6

$$\alpha = \text{Softmax}(WZ_{\text{concat}} + b) \quad (6)$$

The final fused representation Z_{fused} is computed as a weighted sum of the intermediate outputs:

$$Z_{fused} = \sum_{i=1}^{4} \alpha_i \cdot \tilde{Z}_i \quad (7)$$

3.4. Prediction Module

After multimodal fusion, the model obtains a unified semantic representation vector Z_{fused}, which is passed through a multilayer perceptron (MLP) for nonlinear transformation, followed by a sigmoid activation function to produce the predicted probability $\hat{y}$.

During training, we adopt Binary Cross-Entropy Loss with logits as the objective function, defined in Eq. 8:

$$\mathcal{L} = -[y \cdot log(\hat{y}) + (1 - y) \cdot log(1 - \hat{y})] \quad (8)$$

Here, $y \in \{0, 1\}$ is the ground-truth label, and $\hat{y}$ is the predicted probability output by the model.

4. Experiments and Model Implementation

4.1. Datasets

We conduct experiments on two publicly available benchmark datasets: MAMI[30] and CrisisHateMM[31].

The MAMI dataset consists of 11,000 meme-style samples, each comprising an image and its corresponding extracted text. Each sample is annotated with one of five fine-grained categories: non-misogynistic, shaming, stereotype, objectification, or violence. In this study, we focus on the binary classification task of distinguishing misogynistic from non-misogynistic content.

The CrisisHateMM dataset includes 4,723 multimodal samples, each composed of an image and an accompanying text segment. The dataset is annotated for binary hate speech classification: hateful vs. non-hateful. Hateful samples are further divided into targeted and untargeted hate, with targeted instances additionally labeled by the nature of the target group—individual, community, or organization. Notably, this dataset originates from the CASE 2024 Shared Task on Multimodal Hate Event Detection, where the test set labels remain undisclosed. Evaluation is performed through the official competition platform, which provides aggregated performance scores based on submitted predictions.

Both datasets define a binary classification subtask that aligns with the objective of this study: determining whether a given multimodal input expresses hate speech. We adhere to the original train/validation/test splits provided in the official dataset releases. Table 1 summarizes the distribution of samples across these splits.

Dataset	Class	Train	Eval	Test
MAMI	Hate	4973	44	500
	No Hate	4987	56	500
CrisisHateMM	Hate	1942	243	243
	No Hate	1658	200	200

Table 1: Dataset splits for MAMI and CrisisHateMM

4.2. Data Preprocessing and Augmentation

For textual data, we performed basic preprocessing using regular expressions to remove URLs, user mentions, and emojis, retaining only lowercase alphabetic characters. During training, we applied rule-based data augmentation techniques, including token dropout, local token shuffling, and token masking. The processed texts were then tokenized using the RoBERTa tokenizer and padded to a fixed sequence length for input into the model.

For image data, we applied standard augmentation techniques using the Albumentations library, including random cropping, horizontal flipping, affine transformations (e.g., rotation, translation, scaling), and color jittering (brightness, contrast, saturation). All images were subsequently normalized and converted to tensor format. During validation and testing, only resizing and normalization were applied to ensure consistency in evaluation.

4.3. Baseline Models

We compare our proposed MLCA framework with a set of representative baseline models, including eight unimodal and five multimodal approaches. The unimodal baselines consist of four textual encoders: BERT[32], RoBERTa[33], ALBERT[34], and DistilBERT[35]. In addition, we include four vision models: Inception v3[36], ResNet-152[37], DenseNet-161[38], and Swin Transformer V2[28]. These models serve as performance references for scenarios where only a single modality is available.

For multimodal baselines, we evaluate VisualBERT[39], CLIP[40], ViLT[41], FLAVA[42], and BLIP2[43], which span a variety of fusion paradigms, including early fusion, contrastive learning, and unified vision–language modeling. These models represent the current state of the art in multimodal understanding and provide a strong benchmark for evaluating the effectiveness of our proposed method.

To ensure fair comparison, all models were fine-tuned under identical training configurations. On the CrisisHateMM dataset, we evaluated model performance using accuracy, precision, recall, and F1-score. For the MAMI dataset, we report accuracy, F1 score, and AUC score, which is particularly informative in the presence of class imbalance. For both datasets, we adopt macro-averaged F1 to mitigate the impact of label imbalance and provide a more balanced evaluation across classes.

4.4. Experimental Settings

During training, we employed a group-specific learning rate strategy to control the update pace across model components. Learning rates were set to 1e-5 for both the text and image encoders, and 1e-4 for the classification head. We used the AdamW optimizer, which combines adaptive gradient updates with weight decay regularization. A cosine learning rate scheduler with warm-up was applied, where the first 5% of iterations were allocated for warm-up to stabilize early optimization. A batch size of 8 was used. For the CrisisHateMM dataset, models were trained for a maximum of 15 epochs, while for MAMI, training was extended to 20 epochs. In both cases, the model achieving the best macro-F1 score on the validation set was saved for final evaluation. To prevent overfitting and training stagnation, we adopted an early stopping mechanism: training was terminated if the macro-F1 score did not improve for 5 consecutive validation epochs.

Modality	**Model**	**CrisisHateMM**				**MAMI**		
		Acc	**Pre**	**Recall**	**F1**	**Acc**	**F1**	**AUC**
Unimodal-Textual	BERT	0.8239	0.8221	0.8227	0.8224	0.6721	0.6702	0.7422
	RoBERTa	0.8330	0.8331	0.8362	0.8326	0.7065	0.7058	0.7665
	ALBERT	0.7178	0.7177	0.7198	0.7171	0.6684	0.6672	0.7228
	DistilBERT	0.5869	0.5931	0.5929	0.5869	0.6603	0.6586	0.7200
Unimodal-Image	Inception v3	0.6817	0.6787	0.6749	0.6758	0.6337	0.6227	0.6985
	ResNet152	0.6704	0.6676	0.6682	0.6679	0.5872	0.5837	0.6159
	DenseNet	0.6456	0.6425	0.6429	0.6429	0.5989	0.5880	0.6557
	Swin V2	0.7359	0.7367	0.7389	0.7355	0.6390	0.6359	0.7031
Multimodal	VisualBERT	0.7878	0.7939	0.7946	0.7878	0.6798	0.6748	0.6951
	ViLT	0.7494	0.7470	0.7464	0.7467	0.6524	0.6513	0.6951
	CLIP	0.7788	0.7767	0.7762	0.7765	0.7142	0.7102	0.7845
	BLIP-2	0.8126	0.8121	0.8151	0.8121	0.4973	0.4462	0.4576
	FLAVA	0.7652	0.7633	0.7648	0.7638	0.6565	0.6557	0.7352
	MLCA (Ours)	**0.8939**	**0.8942**	**0.8913**	**0.8925**	**0.7564**	**0.7559**	**0.8142**

Table 2: Comparative Performance of Different Models on the CrisisHateMM and MAMI Datasets

5. Results and Discussion

5.1. Evaluation Results

Table 2 presents a comparative performance analysis of our proposed model (MLCA) against a variety of baseline methods on the CrisisHateMM and MAMI datasets. Several key observations can be drawn from the experimental results.

Among unimodal models, textual features contribute more substantially to hate speech detection than visual features. RoBERTa achieves the highest performance, with 83.30% accuracy and 83.26% F1-score on the CrisisHateMM dataset—substantially outperforming all image-only baselines and even surpassing several multimodal approaches. In contrast, smaller models such as ALBERT and DistilBERT yield noticeably lower scores, indicating that model capacity and the depth of pretraining remain critical factors for effective textual modeling.

On the vision side, image-only models consistently underperform relative to their text-based counterparts. The best-performing vision model, Swin Transformer V2, reaches only 73.55% F1-score on CrisisHateMM. This outcome highlights the limited discriminative power of visual features when used in isolation, especially in hate memes that lack overt visual cues, thereby making standalone image-based understanding inherently more challenging.

Multimodal models mitigate the limitations of individual modalities by jointly modeling textual and visual features. For instance, BLIP-2 achieves an F1-score of 81.21% on the CrisisHateMM dataset, approaching the performance of RoBERTa and underscoring the potential of multimodal learning. However, several models—such as VisualBERT, CLIP, ViLT, and FLAVA—still underperform compared to the strongest unimodal text baseline.

Our proposed model, MLCA, achieves the best overall performance on both datasets—obtaining an F1-score of 89.25% on CrisisHateMM and 75.59% on MAMI—substantially outperforming all baseline models. These findings demonstrate the effectiveness of multi-level cross-modal attention and gated fusion in capturing nuanced multimodal hate cues, while also underscoring the critical importance of well-designed fusion architectures in enhancing model performance on this task.

5.2. Comparison with State-of-the-Art

We further compare our model with several recent state-of-the-art (SOTA) methods reported on the CrisisHateMM and MAMI datasets. The comparison results are summarized in Tables 3 and 4.

Model	**Acc**	**Pre**	**Recall**	**F1**
YYama[44]	0.7585	0.7588	0.7613	0.7580
MasonPerplexity[45]	0.8352	0.8347	0.8378	0.8347
ARC-NLP[46]	0.8490	0.8410	0.8900	0.8480
AAST-NLP[47]	–	0.8550	0.8539	0.8544
CLTL[48]	0.8736	0.8720	0.8737	0.8727
MLCA	**0.8939**	**0.8942**	**0.8913**	**0.8925**

Table 3: Performance Comparison with State-of-the-Art Models on the CrisisHateMM Dataset

Model	**Acc**	**F1**	**AUC**
PromptHate[49]	0.7031	–	0.7995
Pro-CapPromptHate[24]	0.7363	–	0.8377
HyperHatePrompt[26]	0.7530	0.7510	**0.8430**
MLCA	**0.7563**	**0.7559**	0.8142

Table 4: Performance Comparison with State-of-the-Art Models on the MAMI Dataset

Extensive experiments on two benchmark datasets, CrisisHateMM and MAMI, demonstrate the effectiveness and robustness of the proposed MLCA model. On CrisisHateMM, prior work explored a wide range of strategies. YYama[44] leveraged prompt-based zero-shot learning with large vision–language models such as LLaVA-1.5B, achieving an F1-score of 75.8%. MasonPerplexity[45] evaluated multiple text encoders, with BERTweet-large reaching 83.47% F1. ARC-NLP[46] combined ELECTRA and Swin Transformer with additional linguistic features, achieving 84.80% F1. AAST-NLP[47] adopted a multi-stage fusion and ensemble strategy, reaching 85.44% F1, while CLTL[48] employed MLP-based fusion and reported the previous best result of 87.27% F1. In comparison, MLCA achieves an F1-score of 89.25%, setting a new state-of-the-art and highlighting the advantage of multi-level cross-modal attention and adaptive fusion in capturing complex and nuanced hate semantics.

On the MAMI dataset, similar trends are observed. Prompt-based models such as PromptHate[49], Pro-CapPromptHate[24], and HyperHatePrompt[26] progressively enhance performance through improved text prompting, image captioning, and cross-modal reasoning. The best prior result was achieved by HyperHatePrompt, with an AUC of 84.30% and an F1-score of 75.10%. In comparison, MLCA slightly outperforms this with an F1-score of 75.59% and a competitive AUC of 81.42%, suggesting strong generalization across both datasets and task configurations.

5.3. Ablation Studies

We conducted ablation studies to evaluate the contributions of four key components: data augmentation, textual self-attention, multi-level cross-modal attention, and gated fusion. As shown in Table 5, removing data augmentation (w/o Augment) led to a noticeable drop in F1-score on both CrisisHateMM (−2.23%) and MAMI (−1.58%), highlighting its role in improving generalization. Eliminating the textual self-attention module (w/o Text-SA) caused moderate performance degradation, confirming its importance in modeling global semantic structure. When the multi-level cross-modal attention mechanism was replaced with a single-layer interaction (w/o Multi-LVL), the model experienced a larger decline in performance, particularly on MAMI (−2.89%), indicating the necessity of hierarchical semantic alignment across modalities. Lastly, removing the gated fusion strategy (w/o Gating) and using mean pooling instead reduced performance on both datasets, which underscores the value of adaptive fusion in effectively integrating multi-level representations.

Models	**CrisisHateMM**				**MAMI**		
	Acc	**Pre**	**Recall**	**F1**	**Acc**	**F1**	**AUC**
MLCA	**0.8939**	**0.8942**	**0.8913**	**0.8925**	**0.7563**	**0.7559**	**0.8142**
-w/o Augment	0.8713	0.8670	0.8707	0.8702	0.7412	0.7401	0.8138
-w/o Text-SA	0.8849	0.8846	0.8827	0.8835	0.7435	0.7429	0.8140
-w/o Multi-Level Fusion	0.8736	0.8722	0.8728	0.8725	0.7311	0.7270	0.7974
-w/o Gating	0.8803	0.8793	0.8790	0.8792	0.7336	0.7322	0.8026

Table 5: Ablation Study Results on the CrisisHateMM and MAMI Datasets

6. Conclusion

In this work, we propose MLCA, a multimodal hate speech detection framework that integrates multi-level cross-modal attention with a gated fusion strategy. By aligning global textual semantics with multi-scale visual features and adaptively aggregating cross-modal interactions across layers, MLCA effectively captures both fine-grained and high-level semantic cues.

Extensive experiments on the CrisisHateMM and MAMI datasets demonstrate that MLCA achieves state-of-the-art performance, surpassing a broad range of unimodal and multimodal baselines. Ablation studies further confirm the essential contributions of multi-level attention, gated fusion, and global text modeling to nuanced hate speech understanding. While multimodal learning enhances model robustness, our analysis reveals that textual signals remain the dominant contributor to performance, whereas visual features offer complementary but less discriminative cues.

These findings underscore the importance of well-structured fusion architectures in effectively leveraging heterogeneous modalities for hate speech detection, and provide practical guidance for the design of future multimodal systems in this domain. While these results are promising, certain limitations suggest directions for future research.

First, the current visual encoder has limited ability to capture implicit or abstract hate signals in complex or context-rich images. Future work could incorporate advanced vision-language pretraining models or visual grounding techniques to enhance visual semantic understanding. Second, while the proposed gated fusion strategy improves modality integration, it may still introduce redundant or noisy representations under semantically sparse or ambiguous conditions.

Exploring adaptive, noise-resilient fusion mechanisms—such as uncertainty modeling or sparse attention—could mitigate this issue and further strengthen model robustness. Third, the framework lacks explicit reasoning components to handle subtle inter-modal dependencies or borderline cases. Introducing lightweight reasoning modules or structured knowledge integration could improve interpretability and decision accuracy.

Author Contributions

This research was primarily conducted by the first author, Rui Lv, and the corresponding author, Lirong Chen. Rui Lv was responsible for formulating the research problem, designing the framework, implementing the model, conducting data collection and analysis, and drafting the manuscript. Lirong Chen provided overall supervision, refined the research objectives, critically revised the manuscript, and secured funding for the project. Jie Wang contributed to data curation and validation of the experiments. Xuan Liu assisted in the analysis and interpretation of the experimental results. All authors have read and approved the final version of the manuscript.

Acknowledgements

This work was supported in part by the National Natural Science Foundation of China under Grant No. 71862027, and in part by the Natural Science Foundation of Inner Mongolia Autonomous Region under Grant No. 2025MS07016.

References

[1] Sushila Shelke and Vahida Attar. Source detection of rumor in social network–a review. *Online Social Networks and Media*, 9:30–42, 2019.

[2] Flor Miriam Plaza-del Arco, M Dolores Molina-González, L Alfonso Urena-López, and M Teresa Martín-Valdivia. Comparing pre-trained language models for spanish hate speech detection. *Expert Systems with Applications*, 166:114120, 2021.

[3] Alexander Tsesis. *Destructive messages: How hate speech paves the way for harmful social movements*, volume 27. NYU Press, 2002.

[4] Yonatan Lupu, Richard Sear, Nicolas Velásquez, Rhys Leahy, Nicholas Johnson Restrepo, Beth Goldberg, and Neil F Johnson. Offline events and online hate. *PLoS one*, 18(1):e0278511, 2023.

[5] Greeshma Arya, Mohammad Kamrul Hasan, Ashish Bagwari, Nurhizam Safie, Shayla Islam, Fatima Rayan Awad Ahmed, Aaishani De, Muhammad Attique Khan, and Taher M Ghazal. Multimodal hate speech detection in memes using contrastive language-image pre-training. *IEEE Access*, 12:22359–22375, 2024.

[6] Sarah Masud, Manjot Bedi, Mohammad Aflah Khan, Md Shad Akhtar, and Tanmoy Chakraborty. Proactively reducing the hate intensity of online posts via hate speech normalization. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 3524–3534, 2022.

[7] Gaurav Rajput, Narinder Singh Punn, Sanjay Kumar Sonbhadra, and Sonali Agarwal. Hate speech detection using static bert embeddings. In *Big Data Analytics: 9th International Conference, BDA 2021, Virtual Event, December 15-18, 2021, Proceedings 9*, pages 67–77. Springer, 2021.

[8] Anna Schmidt and Michael Wiegand. A survey on hate speech detection using natural language processing. In *Proceedings of the fifth international workshop on natural language processing for social media*, pages 1–10, 2017.

[9] Paula Fortuna, Juan Soler-Company, and Leo Wanner. How well do hate speech, toxicity, abusive and offensive language classification models generalize across datasets? *Information Processing & Management*, 58(3):102524, 2021.

[10] Binny Mathew, Punyajoy Saha, Seid Muhie Yimam, Chris Biemann, Pawan Goyal, and Animesh Mukherjee. Hatexplain: A benchmark dataset for explainable hate speech detection. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 14867–14875, 2021.

[11] Sarah Masud, Mohammad Aflah Khan, Md Shad Akhtar, and Tanmoy Chakraborty. Overview of the hasoc subtrack at fire 2023: Identification of tokens contributing to explicit hate in english by span detection. *arXiv preprint arXiv:2311.09834*, 2023.

[12] Maarten Sap, Saadia Gabriel, Lianhui Qin, Dan Jurafsky, Noah A Smith, and Yejin Choi. Social bias frames: Reasoning about social and power implications of language. *arXiv preprint arXiv:1911.03891*, 2019.

[13] Waqas Sharif, Saima Abdullah, Saman Iftikhar, Daniah Al-Madani, and Shahzad Mumtaz. Enhancing hate speech detection in the digital age: A novel model fusion approach leveraging a comprehensive dataset. *IEEE Access*, 2024.

[14] Mai ElSherief, Caleb Ziems, David Muchlinski, Vaishnavi Anupindi, Jordyn Seybolt, Munmun De Choudhury, and Diyi Yang. Latent hatred: A benchmark for understanding implicit hate speech. *arXiv preprint arXiv:2109.05322*, 2021.

[15] Youngwook Kim, Shinwoo Park, and Yo-Sub Han. Generalizable implicit hate speech detection using contrastive learning. In *Proceedings of the 29th international conference on computational linguistics*, pages 6667–6679, 2022.

[16] Sreyan Ghosh, Manan Suri, Purva Chiniya, Utkarsh Tyagi, Sonal Kumar, and Dinesh Manocha. Cosyn: Detecting implicit hate speech in online conversations using a context synergized hyperbolic network. *arXiv preprint arXiv:2303.03387*, 2023.

[17] Junyu Lu, Hongfei Lin, Xiaokun Zhang, Zhaoqing Li, Tongyue Zhang, Linlin Zong, Fenglong Ma, and Bo Xu. Hate speech detection via dual contrastive learning. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 31:2787–2795, 2023.

[18] R Prabhu and V Seethalakshmi. A comprehensive framework for multi-modal hate speech detection in social media using deep learning. *Scientific Reports*, 15(1):13020, 2025.

[19] Douwe Kiela, Hamed Firooz, Aravind Mohan, Vedanuj Goswami, Amanpreet Singh, Pratik Ringshia, and Davide Testuggine. The hateful memes challenge: Detecting hate speech in multimodal memes. *Advances in neural information processing systems*, 33:2611–2624, 2020.

[20] Raul Gomez, Jaume Gibert, Lluis Gomez, and Dimosthenis Karatzas. Exploring hate speech detection in multimodal publications. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 1470–1478, 2020.

[21] Homa Hosseinmardi, Rahat Ibn Rafiq, Richard Han, Qin Lv, and Shivakant Mishra. Prediction of cyberbullying incidents in a media-based social network. In *2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, pages 186–192. IEEE, 2016.

[22] Krishanu Maity, Prince Jha, Sriparna Saha, and Pushpak Bhattacharyya. A multitask framework for sentiment, emotion and sarcasm aware cyberbullying detection from multi-modal code-mixed memes. In *Proceedings of the 45th international ACM SIGIR conference on research and development in information retrieval*, pages 1739–1749, 2022.

[23] Roy Ka-Wei Lee, Rui Cao, Ziqing Fan, Jing Jiang, and Wen-Haw Chong. Disentangling hate in online memes. In *Proceedings of the 29th ACM international conference on multimedia*, pages 5138–5147, 2021.

[24] Rui Cao, Ming Shan Hee, Adriel Kuek, Wen-Haw Chong, Roy Ka-Wei Lee, and Jing Jiang. Pro-cap: Leveraging a frozen vision-language model for hateful meme detection. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 5244–5252, 2023.

[25] Eniafe Festus Ayetiran and Özlem Özgöbek. An inter-modal attention-based deep learning framework using unified modality for multimodal fake news, hate speech and offensive language detection. *Information Systems*, 123:102378, 2024.

[26] Bo Xu, Erchen Yu, Jiahui Zhou, Hongfei Lin, and Linlin Zong. Hyperhateprompt: A hypergraph-based prompting fusion model for multimodal hate detection. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 3825–3835, 2025.

[27] Daniel Loureiro, Kiamehr Rezaee, Talayeh Riahi, Francesco Barbieri, Leonardo Neves, Luis Espinosa Anke, and Jose Camacho-Collados. Tweet insights: a visualization platform to extract temporal insights from twitter. *arXiv preprint arXiv:2308.02142*, 2023.

[28] Ze Liu, Han Hu, Yutong Lin, Zhuliang Yao, Zhenda Xie, Yixuan Wei, Jia Ning, Yue Cao, Zheng Zhang, Li Dong, et al. Swin transformer v2: Scaling up capacity and resolution. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12009–12019, 2022.

[29] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022, 2021.

[30] Elisabetta Fersini, Francesca Gasparini, Giulia Rizzi, Aurora Saibene, Berta Chulvi, Paolo Rosso, Alyssa Lees, and Jeffrey Sorensen. Semeval-2022 task 5: Multimedia automatic misogyny identification. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, pages 533–549, 2022.

[31] Aashish Bhandari, Siddhant B Shah, Surendrabikram Thapa, Usman Naseem, and Mehwish Nasim. 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, 2023.

[32] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*,

pages 4171–4186, 2019.

[33] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.

[34] Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. Albert: A lite bert for self-supervised learning of language representations. *arXiv preprint arXiv:1909.11942*, 2019.

[35] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*, 2019.

[36] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2818–2826, 2016.

[37] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[38] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708, 2017.

[39] Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. Visualbert: A simple and performant baseline for vision and language. *arXiv preprint arXiv:1908.03557*, 2019.

[40] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR, 2021.

[41] Wonjae Kim, Bokyung Son, and Ildoo Kim. Vilt: Vision-and-language transformer without convolution or region supervision. In *International conference on machine learning*, pages 5583–5594. PMLR, 2021.

[42] Amanpreet Singh, Ronghang Hu, Vedanuj Goswami, Guillaume Couairon, Wojciech Galuba, Marcus Rohrbach, and Douwe Kiela. Flava: A foundational language and vision alignment model. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 15638–15650, 2022.

[43] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*, pages 12888–12900. PMLR, 2022.

[44] Yosuke Yamagishi. Yyama@ multimodal hate speech event detection 2024: Simpler prompts, better results-enhancing zero-shot detection with a large multimodal model. In *Proceedings of the 7th workshop on challenges and applications of automated extraction of socio-political events from text (CASE 2024)*, pages 60–66, 2024.

[45] Amrita Ganguly, Al Nahian Bin Emran, Sadiya Sayara Chowdhury Puspo, Md Nishat Raihan, Dhiman Goswami, and Marcos Zampieri. Masonperplexity at multimodal hate speech event detection 2024: Hate speech and target detection using transformer ensembles. *arXiv preprint arXiv:2402.01967*, 2024.

[46] Umitcan Sahin, Izzet Emre Kucukkaya, Oguzhan Ozcelik, and Cagri Toraman. Arc-nlp at multimodal hate speech event detection 2023: Multimodal methods boosted by ensemble learning, syntactical and entity features. *arXiv preprint arXiv:2307.13829*, 2023.

[47] Ahmed El-Sayed and Omar Nasr. Aast-nlp at multimodal hate speech event detection 2024: A multimodal approach for classification of text-embedded images based on clip and bert-based models. In *Proceedings of the 7th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text (CASE 2024)*, pages 139–144, 2024.

[48] Yeshan Wang and Ilia Markov. Cltl@ multimodal hate speech event detection 2024: The winning approach to detecting multimodal hate speech and its targets. In *Proceedings of the 7th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text (CASE 2024)*, pages 73–78, 2024.

[49] Rui Cao, Roy Ka-Wei Lee, Wen-Haw Chong, and Jing Jiang. Prompting for multimodal hateful meme classification. *arXiv preprint arXiv:2302.04156*, 2023.

Author Biography

Rui Lv, a master's student at the College of Computer Science, Inner Mongolia University. His research interests focus on natural language processing, especially multimodal hate speech classification involving both images and text.

Jie Wang, a master's student in the School of Computer Science, Inner Mongolia University. His research interests focus on various aspects of natural language processing, especially in the area of hate speech identification.

Xuan Liu, a master's student at the School of Computer Science, Inner Mongolia University. Her research mainly focuses on natural language processing, in particular on hate speech detection.

Lirong Chen, Associate Professor at the School of Computer Science, Inner Mongolia University. She graduated with a Ph.D. from Dalian University of Technology. Her research interests primarily include e-commerce reputation and trust; fake information detection and hate speech detection on social media platforms; the application of large language models (LLMs) in cross-border e-commerce, etc. She has published multiple high-impact papers in the fields of Natural Language Processing (NLP) and Information Systems.