HyperHatePrompt: A Hypergraph-based Prompting Fusion Model for Multimodal Hate Detection

Bo Xu[1] **Erchen Yu**[1] **Jiahui Zhou**[2] **Hongfei Lin**[1*] **Linlin Zong**[2]
[1] School of Computer Science and Technology, Dalian University of Technology
[2] School of Software, Dalian University of Technology
{xubo, hflin, llzong}@dlut.edu.cn {yuerchen0809, zjhjixiang}@mail.dlut.edu.cn

Abstract

Multimodal hate detection aims to identify hate content across multiple modalities for promoting a harmonious online environment. Despite promising progress, three critical challenges, *the absence of implicit hateful cues*, *the cross-modal-induced hate*, and *the diversity of hate target groups*, inherent in the multimodal hate detection task, have been overlooked. To address these challenges, we propose a hypergraph-based prompting fusion model. Our model first uses tailored prompts to infer implicit hateful cues. It then introduces hyperedges to capture cross-modal-induced hate and applies a diversity-oriented hyperedge expansion strategy to account for different hate target groups. Finally, hypergraph convolution fuses diverse hateful cues, enhancing the exploration of cross-modal hate and targeting specific groups. Experimental results on two benchmark datasets show that our model achieves state-of-the-art performance in multimodal hate detection.

Disclaimer*: The samples presented by this paper may be considered offensive or vulgar.*

1 Introduction

The rapid growth of online communication has facilitated information sharing, enabling individuals from diverse backgrounds to interact with each other. However, the anonymity of the Internet also allows users to express themselves irresponsibly and attack others, leading to a rise in hate content (Kowalski and Whittaker, 2015). Hate content, which includes aggressive, discriminatory and derogatory text and visuals aimed at specific groups based on race, gender, and religion, is a harmful form of online abuse (Jones, 2020). This creates challenges in maintaining a safe and inclusive online space. Multimodal hate content detection, which refers to the process of identifying and analyzing hate-related information presented through multiple modalities (Schmidt and Wiegand, 2017), is of greater significance as it can integrate diverse information from text, image and other modalities, while single-modal detection is limited in capturing comprehensive and accurate cues of hate, thus multimodal approach is essential for a more precise and in-depth understanding and detection of hate content. Therefore, recent research has increasingly focused on the detection of multimodal hate content (Fortuna et al., 2021; Karim et al., 2021; Rajput et al., 2021; Masud et al., 2022; Lu et al., 2023).

The oratory ignites a wave of snoring… **(a)** **Cali…in full RETARD mode** **(b)**

Figure 1: Two examples of multimodal hate detection.

Multimodal hate detection aims to identify hate content across multiple modalities in order to combat online hatred and promote a harmonious online environment. This task has attracted considerable attention in recent years, leading to the successive proposal of various detection models (Botelho et al., 2021; Yang et al., 2022; Hee et al., 2023; Lin et al., 2023). Despite recent progress, three key challenges in multimodal hate detection remain largely overlooked: *the absence of implicit hateful cues, cross-modal-induced hate, and the diversity of hate target groups.*

While previous studies have successfully identified explicit hate across various modalities(Schmidt and Wiegand, 2017), they often miss **the implicit hateful cues**, exemplified by the sarcasm expressions shown in Exp. (a) of Figure 1. Detecting these implicit hateful cues is essential for ac-

*Corresponding Author

Proceedings of the 31st International Conference on Computational Linguistics, pages 3825–3835
January 19–24, 2025.

curately identifying hate speech across multiple modalities. Therefore, it's crucial to equip detection models with the ability to capture these hidden meanings in multimodal hate content.

Although advancements have been made in combining multimodal hate information(Wiegand et al., 2019), the phenomenon of **the cross-modal-induced hate** remains relatively unexplored. The cross-modal-induced hate refers to the content that doesn't possess hate characteristics in each single modality but exhibits hate features when combined together. In Exp. (b) of Figure 1, hate emerges from the interaction between content across modalities, even when individual pieces do not display obvious hate. This highlights the need for more research into cross-modal fusion to fully understand the hate content. Compared with implicit hate conveyed through implicit emotions, cross-modal hate places greater emphasis on the interactive modeling among modalities in order to achieve better detection performance.

A third challenge lies in the lack of modeling **the diversity of hate target groups**. Current models often overlook the different backgrounds, languages, and perspectives of various groups affected by hate content (Charitidis et al., 2020). This lack of representation weakens the fairness and inclusivity of hate detection, limiting their ability to capture the full context of hate across diverse demographics. Addressing this diversity is essential for improving the performance and fairness of multimodal hate detection.

To address these challenges, we propose a hypergraph-based prompting fusion model, HyperHatePrompt, for multimodal hate detection. To capture *the implicit hateful cues*, we design an implicit hate cue prompt that infers hate semantics beyond the textual domain. The prompted cues are then treated as a unique modality in multimodal learning, providing in-depth understanding of hate. To fully comprehend *the cross-modal-induced hate*, we introduce hypergraphs to capture hate-related aspects across modalities by constructing high-order hyperedges. Unlike regular graphs that connect only two nodes, hyperedges connect multiple nodes, conveying diverse hateful cues from different modalities, thus enabling a more comprehensive learning of cross-modal-induced hate (Xu et al., 2023). While concatenating cross-modal features, as in previous works (Hee et al., 2023; la Peña Sarracén, 2021; Botelho et al., 2021), can achieve a certain level of cross-modal ability, hypergraphs offer a more sophisticated and comprehensive way to model the complex relationships among different modalities. Hyperedges can connect multiple nodes from different modalities simultaneously, which allows for a more fine-grained and accurate representation of the interactions and dependencies between various types of hate cues. To consider *the diversity of hate target groups* in multimodal learning, we propose a novel diversity-oriented hyperedge expansion strategy, which updates the hypergraph based on the diversity divergence between hyperedges. Through hypergraph convolution, diversified hateful cues across modalities are fused, distinguishing node features for hate samples targeting specific groups and enhancing the exploration of cross-modal-induced hate. The main contributions of our work are summarized as follows:

- We explore implicite hateful cues, cross-modal-induced hate and diverse hate target groups in multimodal hate detection, offering new perspectives and deeper insights into the detection process.
- We propose a novel hate detection model that leverages LLM-driven prompting and hypergraph learning with a customized hyperedge expansion strategy to fully capture the intricate semantics in multimodal hate content.
- We evaluate our model on two benchmark datasets, and demonstrate its superiority over state-of-the-art baselines in multimodal hate detection through extensive experiments.

2 Related Work

Our work primarily concerns two lines of related work: single-modal and multimodal hate detection.

2.1 Single-modal Hate Detection

In early single-modal hate detection, the focus was mainly on analyzing text, leading to the creation of datasets like HateXplain (Mathew et al., 2021) and USElectionHate (Grimminger and Klinger, 2021). Researchers then used pre-trained language models for detecting hate and improved methods with various techniques: Fortuna et al. (2021) used different pre-trained language models for detection. Karim et al. (2021) applied multiple models to detect Bengali hate speech. Rajput et al. (2021) enhanced contextual understanding by combining deep neural networks with BERT embeddings. Masud et al.

(2022) created a parallel corpus of hate speech and normalized versions to reduce hate speech severity. Clarke et al. (2023) introduced a exemplar-based, explainable learning approach for hate speech detection. Ocampo et al. (2023) categorized implicit hate messages by complexity levels.

As research has progressed, traditional text-based methods are not enough for the complexities of multimodal social media. Thus, incorporating information from various modalities is becoming crucial for effective hate detection.

2.2 Multimodal Hate Detection

Multimodal hate detection has gained significant attention, especially following initiatives like Facebook's hate memes challenge (Kiela et al., 2020). Recent studies have focused on detecting hate memes across different media types by incorporating both visual and textual information (Sharma et al., 2022; Suryawanshi et al., 2020; Liu et al., 2022; Gasparini et al., 2022; Hossain et al., 2022; Pramanick et al., 2021; Shang et al., 2021; Zhou et al., 2021). Multimodal hate detection broadens the spectrum of hate meme detection to encompass diverse information across modalities (Chhabra and Vishwakarma, 2023; Fersini et al., 2022; Gomez et al., 2020; Hee et al., 2023; Bhandari et al., 2023; Thapa et al., 2022). For instance, Lee et al. (2021) used R-CNN and BERT to improve hate speech detection by identifying key entities. la Peña Sarracén (2021) applied graph convolutional neural networks for multilingual hate detection. Botelho et al. (2021) explored how context helps in detecting both implicit and explicit hate. Yang et al. (2022) used knowledge of irony to improve hate detection. Similarly, Chauhan et al. (2022) used multimodal attention to detect ironic expressions in hate content. Cao et al. (2023) used a pre-trained vision-language model to generate helpful captions for detecting hateful memes. Beyond just detecting hate content, Hee et al. (2023) and Lin et al. (2023) reduce biases in multimodal hate detection models. With the rise of large language models, researchers are focusing on using large vision-language models and creating prompt templates (Cao et al., 2022a).

The above methods overlook the implicit hateful cues, the cross-modal-induced hate, and the diversity of hate target groups. Therefore, we design an hypergraph-based prompting fusion model.

3 Proposed Approach

3.1 Model Overview

The objective of multimodal hate detection is to identify various forms of hate content conveyed through multimodal data, encompassing both textual and visual elements. Specifically, a multimodal hate detection dataset $D = (X, Y)$ consists of pairs of data samples (x_i, y_i), where $x_i \in X$ represents the input multimodal information, and $y_i \in Y$ denotes the ground-truth labels. The input x_i typically consists of a text description t_i and an image m_i, forming the tuple $X = (T, I)$. The learning goal is to identify whether a data sample is hate or not by collectively considering the semantic cues presented in both the text and image modalities, predicting the corresponding hate label y. Therefore, multimodal hate detection models can be regarded as a mapping function $f : T \times I \to y$.

To this end, we propose our HyperHatePrompt model, and illustrate its main architecture in Figure 2. Our model comprises four key modules: implicit hate cue prompting module, hyperedge construction module, hypergraph learning module, and hate label prediction module. The implicit hate cue prompting module utilizes LLMs to prompt implicit hateful cues beyond text. The hyperedge construction module aggregates highly expressive hate-related aspects from text, image, and prompts, and constructs high-order hyperedges across modalities to model cross-modal-induced hate. The hypergraph learning module employs a diversity-oriented hyperedge expansion strategy with hypergraph convolution centering on hate target groups to fuse diverse hateful cues across modalities. The hate label prediction module utilizes the fused hypergraph representations to predict the hate labels of each data sample.

3.2 Implicit Hate Cue Prompting Module

Textual hate speech often manifests implicitly, targeting specific demographic groups with intricate semantics that extend beyond mere negative emotions. This poses a significant challenge in identifying implicit hateful cues embedded within text. To tackle this challenge, we leverage the notable commonsense reasoning abilities of LLMs, and devise a prompt template aimed at revealing implicit hate viewpoints towards certain demographic groups. Our prompting template is shown as follows.

You are a helpful assistant designed to detect

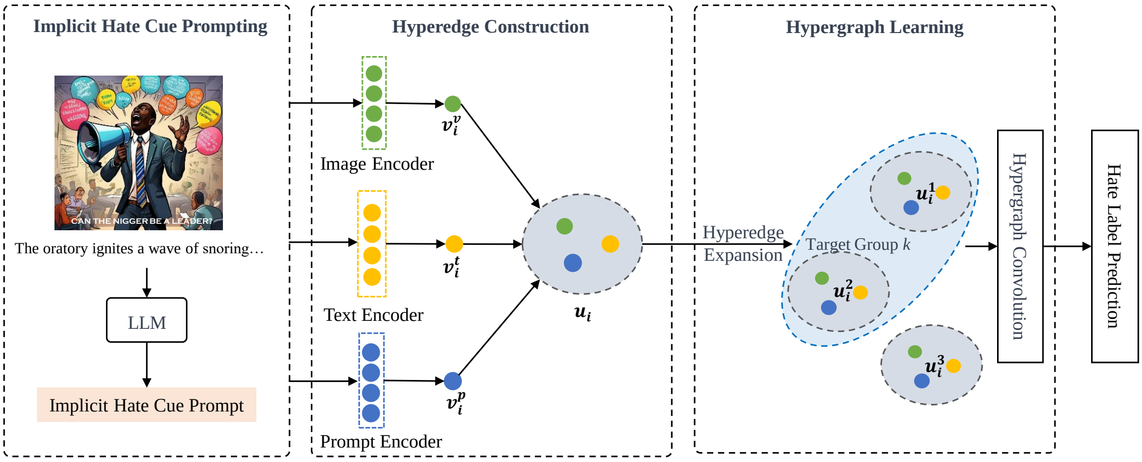


Figure 2: The main architecture of our HyperHatePrompt model.

hate speech. Infer the implicit semantic information of the following text targeted a certain demographic group. Please begin with "the text contains" in your response. Text: {text}.

In our template, LLMs serve as a helpful assistant designed to amplify implicitly expressed hateful cues. In our implementation, we utilize the GPT-3.5-turbo model via OpenAI API, employing zero-shot generation to generate prompts that convey implicit hateful cues.

3.3 Hyperedge Construction Module

To capture hateful cues in each individual modality, we employ the contrastive language–image pretraining model (CLIP) (Radford et al., 2021) as the image and text encoder to derive initial representations of each modality. Namely, we utilize the CLIP-ViT-B-32 model to extract image features, and the Transformer encoder within CLIP to extract text and prompt features, respectively, yielding the image feature representation v_i^m, the text feature representation v_i^t, and the prompt feature representation v_i^p. These three features are served as three types of nodes conveying diverse hateful cues in each modality.

To further consider cross-modal-induced hate, we construct hyperedges through connecting nodes of different modalities within each data sample. The hyperedge u_i of the i-th sample is constructed as $u_i = \{v_i^m, v_i^t, v_i^p\}$, thereby capturing both intrinsic modality-specific features and coupled inter-modal features. We define the initial graph representation V as the concatenation of node representations across different modalities in the same batch, and obtain the representations of hyperedges U as follows.

$$U = B^{-1} \cdot W_e \cdot H^T \cdot V \tag{1}$$

where H is the matrix of associations between modalities, defined as in Eq.(2). B is the degree matrix of hyperedges with $B_{ii} = \sum_j H_{ij}$. W_e is the weight matrix of hyperedges, which is an identity matrix because each hyperedge is assigned equal importance.

$$H(i,j) = \begin{cases} 1, & \text{if node } i \text{ is in hyperedge } j \\ 0, & \text{if node } i \text{ is not in hyperedge } j \end{cases} \tag{2}$$

The coupled inter-modal features in hyperedges focuses on the jointly modeling of cross-modal-induced hate by comprehensively fusion of multimodal hateful cues, facilitating further learning of hypergraphs for hate detection.

3.4 Hypergraph Learning Module

3.4.1 Hyperedge Expansion

The hyperedge construction phase extracts feature representations of each modality, and establishes the initial hyperedges across modalities. Considering the diversity of hate target groups, we propose to enhance the specificity of hate detection across diverse hate target groups by expanding the hyperedges, consolidating those targeting the same groups and discretizing those targeting different groups. To this end, we design a diversity-oriented hyperedge expansion strategy to capture the multidimensional relationships of hate-related aspects across target groups.

Specifically, to prevent over-smoothing of node embeddings, we adopt a breadth-first hyperedge expansion strategy. Given that there is no overlap between our initial hyperedges, there is no inherent connectivity among them. Hence, we utilize the Manhattan distance between hyperedges to assess the diversity divergence for expanding the hyperedge i to the hyperedge j, calculated as follows.

$$\alpha(i,j) = \sum_{k=1}^{d\_e} |U_{ik} - U_{jk}| \tag{3}$$

where U_{ik} is the k-th dimension in the i-th hyperedge representation, and d_e is dimension of hyperedge representation. Following this equation, we parallelly expand all hyperedges. For each hyperedge, we select the top-k hyperedges with the highest diversity likelihood of expansion, and combine them to form a new hyperedge.

After expansion, duplicate hyperedges targeting the same groups are eliminated, and the association matrix H is updated in hyperedge expansion. Notably, unlike the existing hyperedge expansion strategy (Sun et al., 2021), our model retains the initial hyperedges after expansion to ensure continued connectivity between nodes of different modalities within the same sample, thereby facilitating the exploration of cross-modal-induced hate. Through the diversity-oriented hyperedge expansion, our model augments the number of hyperedges across diverse hate target groups, yielding a hierarchical multi-level hyperedge structure. A hierarchical multi-level hyperedge structure involves hyperedges at multiple layers as so to represent complex relationships in a hierarchical way, thus modeling cross-modal information effectively. This expansion strategy enables a deeper understanding of hate semantics, breaking the constraints of pairwise relationships and effectively addressing the diversity of hate target groups.

3.4.2 Hypergraph Convolution

Our model integrates multimodal features via hypergraph convolution centering on hate target groups, calculated as follows.

$$V^{(l+1)} = D^{-1} \cdot H \cdot W_e \cdot B^{-1} \cdot H^T \cdot V^{(l)} \tag{4}$$

where l represents the l-th convolutional layer, starting from 0, W_e is the weight matrix of hyperedges, H is the updated matrix of associations between modalities, $V^{(0)}$ is the initial graph representation in Eq.(1), D is the degree matrix of nodes, and B is the degree matrix of hyperedges. The metrices D and B are diagnoal matrix and updated as the matrix H evolves, with $D_{jj} = \sum_i H_{ij}$ and $B_{ii} = \sum_j H_{ij}$. We design this hypergraph convolution centering on hate target groups to ensure effective detection without relying on non-linear activation and convolutional filters, thus reducing model complexity for enhanced training speed.

This process of hypergraph convolution follows an aggregation order of "node-hyperedge-node", wherein features are aggregated from nodes to hyperedges and then from hyperedges to nodes, thus aggregates multi-level hateful cues across modalities. The multi-level hyperedge structure enables convolution-derived features to encompass information for diverse hate target groups, and enhances the cross-modal-induced hate. For instance, if node x_1^t contained in both the inital hyperedge u_1 and the expanded hyperedge u_2 (where u_1 is a subset of u_2), then x_1^t aggregates information not only from the hyperedge u_1, but also from the hyperedge u_2. Despite all hyperedges having a default equal weight, nodes from the same sample but different modalities undergo more aggregations compared to nodes from different samples. This results in more diversified node features for hate samples targeting specific groups, thereby further enhancing cross-modal-induced hate.

3.5 Hate Label Prediction Module

The final sample representation is determined as follows:

$$P = \frac{1}{(L+1)} \sum_{l=0}^{L} V^{(L)} \tag{5}$$

where L is the number of convolution layers. Based on the fully fused hate feature representations, a multi-layer perceptron is used as the final classifier. Namely, the fused representations are fed into the perceptron to predict the hate label as follows.

$$P' = \tanh(W_1 P + b_1) \tag{6}$$

$$Y = \text{sigmoid}(W_2 P' + b_2) \tag{7}$$

where W_1, b_1, W_2, and b_2 are the parameters of two fully connected layers.

4 Experiments

4.1 Datasets

We conducted experiments on two benchmark datasets: MMHS150K (Gomez et al., 2020) and MAMI (Fersini et al., 2022). The MMHS150K dataset, sourced from Twitter, comprises six hate categories: non-hate, racist, sexist, homophobic, religion-based hate, and other hate tweets, consisting of 149,823 samples in total. Each sample contains both text and image, with some images containing textual content. The MAMI dataset, derived from Semeval-2022 Task 5, focuses on the detection of misogynistic viewpoints, sourced from Twitter, Reddit, and meme-based websites. It includes five misogyny categories: not misogynous, shaming, stereotype, objectification, and violence, consisting of 11,000 samples. Each sample comprises a pair of image and text extracted from the image. We follow the original division of both datasets, as shown in Table 1.

Dataset	Train	Validation	Test
MMHS150K	134,823	5,000	10,000
MAMI	9,500	500	1,000

Table 1: Division of MMHS150K and MAMI datasets.

4.2 Baselines

We compared our model with several baselines, including four single-modal models and five multimodal models. For single-modal models, we compared with BERT (Devlin et al., 2018) and CLIP (Radford et al., 2021) for text-modality modeling, and ResNet (He et al., 2016) and CLIP for image-modality modeling. For multimodal models, we compared with EF-CaTrBERT (Khan and Fu, 2021), CAFE (Chen et al., 2022), TOT (Zhang et al., 2023), PromptHate (Cao et al., 2022b) and Pro-Cap (Cao et al., 2023). EF-CaTrBERT is a dual-stream model for image-text classification, which incorporates images into the auxiliary sentences of the text encoder and feeds them into the model upon fusion. CAFE employs cross-modal fuzzy perception to adaptively aggregate distinctive cross-modal relevant features and single-modal features to reduce mis-classification caused by inter-modality inconsistency. TOT is a topology-aware framework to decipher the implicit harmful memes for optimal transportation plan based cross-modal aligning. PromptHate is a prompt-based model that prompts pre-trained language models for hateful meme classification. Pro-Cap utilizes the frozen pre-trained vision-language model to generate captions that contain information useful for hateful meme detection. To ensure fair comparisons, we fine-tuned all models under identical settings. We evaluated the performance using accuracy, F1 score, and AUC score. Given the potential class imbalance in these datasets, we employed macro-average scores for F1 metric.

Methods	Accuracy	Macro-F1	AUC
BERT-text	0.643	0.642	0.685
CLIP-text	0.677	0.677	0.724
ResNet-image	0.501	0.342	0.534
CLIP-image	0.585	0.585	0.629
CAFE	0.657	0.649	0.691
TOT	0.676	0.674	0.722
EF-CaTrBERT	0.672	0.670	0.711
PromptHate	0.679	0.679	0.729
Pro-Cap	0.712	0.711	0.793
HyperHatePrompt	**0.757**	**0.757**	**0.841**

Table 2: Performance comparisons on MMHS150K.

4.3 Implementation Details

We fine-tuned all model hyperparameters on the validation set. The feature size of CLIP was set to 512. In hypergraph learning module, k in hyperedge expansion was set to 8, and L in hypergraph convolution was set to 3. We employed the Adam optimizer (Kinga and Adam, 2015) with an initial learning rate of 1e-5, and L2 weight decay of 1e-3. The training batch size was set to 64, and the dropout rate was set to 0.5. For the MMHS150K dataset, the classification threshold was set to 0.5, and the number of epochs was set to 20. For the MAMI dataset, the classification threshold was set to 0.8 in consideration of data imbalance, and the number of epochs was set to 100. Training would terminate if the macro-F1 performance on the validation set did not improve within 10 epochs. We have released our code[1] for reproduction.

4.4 Results and Discussions

We present the accuracy, macro-F1 score, and AUC score of our model and baseline models in Table 2 and Table 3. From the results, we observe that:

(1) For single-modal models, BERT-based textual modeling exhibited moderate performance, whereas CLIP-based textual modeling surpassed BERT, notably outperforming all other baseline models on the MMHS150K dataset. Conversely, ResNet-based image modeling demonstrated inferior performance, while CLIP-based image modeling emerged as the top performer among all the baselines on the MAMI dataset. These findings sug-

[1] https://github.com/Meraki2189/HyperHatePrompt

Methods	Accuracy	Macro-F1	AUC
BERT-text	0.576	0.527	0.661
CLIP-text	0.619	0.602	0.703
ResNet-image	0.609	0.587	0.682
CLIP-image	0.705	0.705	0.812
CAFE	0.599	0.578	0.648
TOT	0.658	0.657	0.727
EF-CaTrBERT	0.678	0.672	0.749
PromptHate	0.711	0.708	0.808
Pro-Cap	0.733	0.724	0.832
HyperHatePrompt	**0.753**	**0.751**	**0.843**

Table 3: Performance comparisons on MAMI.

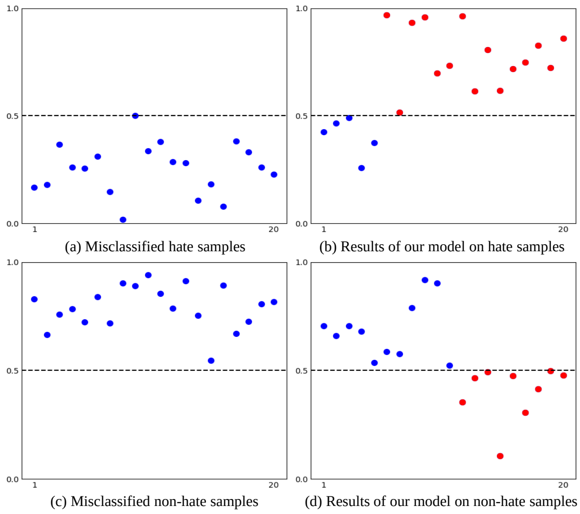


Figure 3: Analysis of misclassified samples by Pro-Cap, where the horizontal axis denotes representations of samples in one-dimensional space, while the vertical axis denotes the probability of each sample being predicted as hate.

gest that the textual data in MMHS150K provide abundant information, facilitating the detection of hate content with rich semantic cues, whereas the images in MAMI similarly provide substantial information for effective hate detection.

(2) For multimodal models, the performance consistency across different models was more notable and considerable. Overall, their performance tended to surpass those of single-modal models, albeit with a slightly weaker best performance. This suggests that current multimodal models face challenges in effectively integrating multimodal hate information, resulting in slightly inferior best performance compared to single-modal counterparts.

(3) Overall, our HyperHatePrompt model achieved the best performance, with improvements of 4.5% in accuracy, 4.6% in macro-F1, and 4.8% in AUC on the MMHS150K dataset compared to the best-performed baseline model. On the MAMI dataset, our model achieved improvements of 2% in accuracy, 2.7% in macro-F1, and 1.1% in AUC compared to the best-performed baseline model. This highlights the effectiveness of hypergraph-based prompting fusion in our model, contributing to better understanding of intricate hate semantics for more accurate hate detection.

4.5 Ablation Study

We conducted ablation studies of our model to investigate the impact of different modalities, hypergraph-based modules, encoders and LLM-based prompting, respectively. The ablated results are shown in Table 4.

(1) **The impact of different modalities**. Removing the representations of each modality from our model resulted in varying degrees of performance degradation across all metrics, indicating that all three modalities contribute to the overall performance. Specifically, removing the textual modality on MMHS150K led to a more significant performance drop, suggesting that textual hate expressions are more prominent in this dataset while removing the image modality on MAMI resulted in a more significant performance drop, indicating that image of hate content is more pronounced in this dataset. Removing prompts on both datasets resulted in a performance decrease, indicating that prompts provide sufficient yet useful hateful cues.

(2) **The impact of hypergraph modules**. Removing the hyperedge expansion strategy and directly classifying the initial hyperedges resulted in decreased model performance. Moreover, removing the hypergraph, we directly concatenated the representations of images, text, and prompts, and fed them into the final classifier, which also led to a significant performance drop on both datasets. These results underscore that the critical role of hypergraphs in detecting cross-modal-induced hate and facilitating diversity-oriented fusion centering on diverse hate target groups. This hypergraph-based learning produces complementary effects, enriching the higher-order semantic information on hateful cues for effective hate detection.

(3) **The impact of encoders and LLMs**. We conducted ablation experiments by replacing the CLIP-based text encoder with the BERT encoder, the CLIP-based image encoder with the ResNet encoder, and ChatGPT with FLAN-T5 for prompting. The results of these ablations revealed varying

Methods	MMHS150K			MAMI		
	Accuracy	**Macro-F1**	**AUC**	**Accuracy**	**Macro-F1**	**AUC**
HyperHatePrompt	0.757	0.757	0.841	0.753	0.751	0.843
- Text Modality	0.719	0.719	0.790	0.741	0.738	0.829
- Image Modality	0.744	0.744	0.835	0.650	0.650	0.703
- Prompt Modality	0.742	0.742	0.830	0.734	0.731	0.824
- Hyperedge Expansion	0.682	0.681	0.739	0.721	0.721	0.826
- Hypergraph Learning	0.676	0.675	0.727	0.715	0.712	0.815
+ BERT Encoder	0.706	0.706	0.771	0.745	0.744	0.832
+ ResNet Encoder	0.709	0.708	0.773	0.717	0.717	0.776
+ FLAN-T5 Prompt	0.750	0.750	0.832	0.745	0.743	0.829

Table 4: Ablation studies on MMHS150K and MAMI.

degrees of performance degradation in our model, although the BERT encoder and FLAN-T5-based prompting yielded comparable performance. These findings highlight the significance of the used encoders and prompting in capturing implicit hateful cues for effectively comprehending hate content in multimodal data.

4.6 Case Study

To illustrate the effectiveness of HyperHatePrompt, we presented case studies in Table 5 and Figure 3. Table 5 illustrates the generated prompts and the predictions made by each model. All these hate cases used the combination of implicit clues and image information for joint detection. From the results, it is observed that our model achieved correct predictions on all three cases, while most baseline models failed on certain cases. This can be attributed to the hateful cues obtained from the prompts (highlighted in blue) and the role of hypergraph learning in integrating multimodal hateful cues.

To conduct a statistical analysis on misclassified cases, we randomly selected twenty hate samples that were misclassified by Pro-Cap, as shown in Exp. (a) of Figure 3. We applied our model to classify these samples, resulting in the correction of fifteen samples shown in Exp. (b). Conversely, we randomly selected twenty non-hate samples misclassified by Pro-Cap in Exp. (c), and applied our model for classification, presented in Exp. (d), which led to the correction of nine samples. Unlike baseline models, which demonstrated inconsistencies in their predictions across these cases, our model consistently achieved superior performance. This outcome can be attributed to the utilization of hyperedge expansion and hypergraph learning techniques, which model the cross-modal-induced hate and address the diversified hate target groups.

5 Conclusions

In this work, we introduce HyperHatePrompt, a novel hypergraph-based model for multimodal hate detection that addresses three key challenges: *implicit hateful cues*, *cross-modal-induced hate*, and *the diversity of hate target groups*. Our model uses LLMs to generate hate cue prompts and applies hypergraph learning with a tailored hyperedge expansion strategy to merge multimodal hate features and enhance the exploration of cross-modal hate and targeting specific groups. Experiments on two benchmark datasets show that HyperHatePrompt outperforms state-of-the-art models. Future research could focus on optimizing prompting strategies and refining multimodal fusion techniques for even better performance.

6 Limitations

While our model shows promise in detecting multimodal hate content targeting different groups, there are some limitations to consider. Its accuracy depends on the quality of the training data, so any biases or gaps in the data can lead to biased or incorrect predictions, especially for underrepresented communities. It may also have difficulty identifying subtle or nuanced hate content that isn't fully captured by the prompts or features. Overcoming these challenges is key to improving multimodal hate detection in real-world settings.

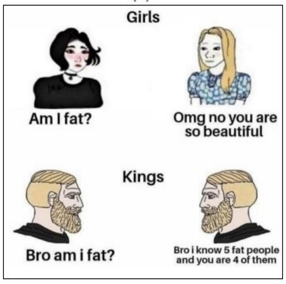
	(a)	(b)	(c)
Hate Case	 Where will you be when diarrhea hits in deep shit by like a boss.	 Out of the five overweight people I know, you're four of them.	 The wonderful singing like a howling.
Prompt	The text contains insensitivity towards individuals experiencing hardship, suggesting a lack of empathy for their well-being.	The text contains derogatory remarks regarding the overweight of a specific individual or demographic group in an ironic undertone.	The text contains a demeaning comparison that the singing is harsh or unpleasant, akin to the sound of a wolf's howling.
Prediction	BERT: ✗, CLIP-text: ✓, ResNet: ✗, CLIP-image: ✗, CAFE: ✗, TOT: ✗, EF-CaTrBERT: ✗, PromptHate: ✗, Pro-Cap: ✓, HyperHatePrompt: ✓.	BERT: ✓, CLIP-text: ✗, ResNet: ✗, CLIP-image: ✗, CAFE: ✗, TOT: ✗, EF-CaTrBERT: ✗, PromptHate: ✗, Pro-Cap: ✗, HyperHatePrompt: ✓.	BERT: ✗, CLIP-text: ✗, ResNet: ✗, CLIP-image: ✗, CAFE: ✓, TOT: ✗, EF-CaTrBERT: ✗, PromptHate: ✗, Pro-Cap: ✗,HyperHatePrompt: ✓.

Table 5: Illustration of case study. Contents highlighted in blue within the prompts are the implications of hateful cues. ✓ and ✗ denote correctly and incorrectly predictions, respectively.

7 Ethics Statement

As researchers in multimodal hate detection, we prioritize ethical use of hate data, focusing on fairness, equity, and inclusivity to reduce biases and protect human rights. We are mindful of the societal impact and aim to combat online hate without causing harm. We value transparency by openly sharing our methods, code, and results, and engage with communities, policymakers, and advocacy groups to ensure our work aligns with ethical standards. Our goal is to conduct research that promotes social cohesion, inclusivity, and justice.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China (No. 62376051), the Fundamental Research Funds for the Central Universities (DUT24MS003), and the Liaoning Provincial Natural Science Foundation Joint Fund Program(2023-MSBA-003).

References

Ayush Bhandari, Saurav B Shah, Samrat Thapa, and et al. 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 1993–2002.

Austin Botelho, Scott A. Hale, and Bertie Vidgen. 2021. Deciphering implicit hate: Evaluating automated detection algorithms for multimodal hate. ACL/IJCNLP 2021:1896–1907.

Rui Cao, Ming Shan Hee, Adriel Kuek, Wen-Haw Chong, Roy Ka-Wei Lee, and Jing Jiang. 2023. 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.

Rui Cao, Roy Ka-Wei Lee, Wen-Haw Chong, and Jing Jiang. 2022a. Prompting for multimodal hateful meme classification. pages 321–332.

Rui Cao, Roy Ka-Wei Lee, Wen-Haw Chong, and Jing Jiang. 2022b. Prompting for multimodal hateful meme classification. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 321–332.

Polychronis Charitidis, Stavros Doropoulos, Stavros Vologiannidis, Ioannis Papastergiou, and Sophia Karak eva. 2020. Towards countering hate speech against journalists on social media. *Online Social Networks and Media*, 17:100071.

Deepak Singh Chauhan, Gurpreet Virdi Singh, Aakash Arora, and et al. 2022. An emoji-aware multitask framework for multimodal sarcasm detection. *Knowledge-Based Systems*, 257:109924.

Yu Chen, Dongsheng Li, Peng Zhang, and et al. 2022. Cross-modal ambiguity learning for multimodal fake news detection. In *Proceedings of the ACM Web Conference 2022*, pages 2897–2905.

Anant Chhabra and Dileep Kumar Vishwakarma. 2023. A literature survey on multimodal and multilingual automatic hate speech identification. *Multimedia Systems*, pages 1–28.

Christopher Clarke, Matthew Hall, Gaurav Mittal, Ye Yu, Sandra Sajeev, Jason Mars, and Mei Chen. 2023. Rule by example: Harnessing logical rules for explainable hate speech detection. In *Proceedings* *of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 364–376. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and et al. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Elisabetta Fersini, Federico Gasparini, Gianluca Rizzi, and et al. 2022. Semeval-2022 task 5: Multimedia automatic misogyny identification. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, pages 533–549.

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

Federico Gasparini, Gianluca Rizzi, Alberto Saibene, and et al. 2022. Benchmark dataset of memes with text transcriptions for automatic detection of multi-modal misogynistic content. *Data in brief*, 44:108526.

Roger Gomez, Jordi Gibert, Lluís Gomez, and et al. 2020. Exploring hate speech detection in multimodal publications. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 1470–1478.

Lara Grimminger and Roman Klinger. 2021. Hate towards the political opponent: A twitter corpus study of the 2020 US elections on the basis of offensive speech and stance detection. pages 171–180.

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

Ming Shan Hee, Wen-Haw Chong, and Roy Ka-Wei Lee. 2023. Decoding the underlying meaning of multimodal hateful memes. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence, IJCAI 2023*, pages 5995–6003. ijcai.org.

Emroz Hossain, Omar Sharif, and Md Mahmudul Hoque. 2022. Mute: A multimodal dataset for detecting hateful memes. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing: Student Research Workshop*, pages 32–39.

Rachel Jones. 2020. Hate speech in the digital age: A review of the literature. *Journal of Policing, Intelligence and Counter Terrorism*, 15(2):103–119.

Md Rabiul Karim, Soumya Kanti Dey, Tareq Islam, and et al. 2021. Deephateexplainer: Explainable hate speech detection in under-resourced bengali language. In *2021 IEEE 8th International Conference on Data Science and Advanced Analytics (DSAA)*, pages 1–10. IEEE.

Zaid Khan and Yun Fu. 2021. Exploiting bert for multimodal target sentiment classification through input space translation. In *Proceedings of the 29th ACM international conference on multimedia*, pages 3034–3042.

Douwe Kiela, Hamed Firooz, Ankur Mohan, and et al. 2020. The hateful memes challenge: Detecting hate speech in multimodal memes. *Advances in neural information processing systems*, 33:2611–2624.

Diederik Kinga and Jimmy Ba Adam. 2015. A method for stochastic optimization. In *International conference on learning representations (ICLR)*, volume 5.

Robin M. Kowalski and Elizabeth Whittaker. 2015. Cyberbullying. In *The Wiley Handbook of Psychology, Technology, and Society*, chapter 8.

Gretel Liz De la Peña Sarracén. 2021. Multilingual and multimodal hate speech analysis in twitter. In *WSDM '21, The Fourteenth ACM International Conference on Web Search and Data Mining*, pages 1109–1110. ACM.

Ryan K W Lee, Rongzheng Cao, Zhiwei Fan, and et al. 2021. Disentangling hate in online memes. In *Proceedings of the 29th ACM international conference on multimedia*, pages 5138–5147.

Hongzhan Lin, Ziyang Luo, Jing Ma, and Long Chen. 2023. Beneath the surface: Unveiling harmful memes with multimodal reasoning distilled from large language models. pages 9114–9128.

Chen Liu, Garrette Geigle, Ryan Krebs, and et al. 2022. Figmemes: A dataset for figurative language identification in politically-opinionated memes. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 7069–7086.

Jinhao Lu, Hongfei Lin, Xing Zhang, and et al. 2023. Hate speech detection via dual contrastive learning. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*.

Sajal Masud, Manik Bedi, Md Atiqul Khan, and et al. 2022. 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.

Binny Mathew, Punyajoy Saha, Seid Muhie Yimam, and et al. 2021. Hatexplain: A benchmark dataset for explainable hate speech detection. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 14867–14875.

Nicolás Benjamín Ocampo, Elena Cabrio, and Serena Villata. 2023. Playing the part of the sharp bully: Generating adversarial examples for implicit hate speech detection. In *Findings of the Association for Computational Linguistics: ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 2758–2772. Association for Computational Linguistics.

Soumyajit Pramanick, Shubham Sharma, Dimitar Dimitrov, and et al. 2021. Momenta: A multimodal framework for detecting harmful memes and their targets. *arXiv preprint arXiv:2109.05184*.

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, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR.

Gaurav Rajput, Narinder Singh Punn, Shweta Kumari Sonbhadra, and et al. 2021. 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 International Publishing.

Anna Schmidt and Michael Wiegand. 2017. 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.

Lu Shang, Yuyu Zhang, Yu Zha, and et al. 2021. Aomd: An analogy-aware approach to offensive meme detection on social media. *Information Processing & Management*, 58(5):102664.

Shubham Sharma, Faisal Alam, Md Shad Akhtar, and et al. 2022. Detecting and understanding harmful memes: A survey. *arXiv preprint arXiv:2205.04274*.

Xiaolong Sun, Hongzhi Yin, Bin Liu, and et al. 2021. Multi-level hyperedge distillation for social linking prediction on sparsely observed networks. In *Proceedings of the Web Conference 2021*, pages 2934–2945.

Shital Suryawanshi, Bharathi Raja Chakravarthi, Mihaela Arcan, and et al. 2020. Multimodal meme dataset (multioff) for identifying offensive content in image and text. In *Proceedings of the second workshop on trolling, aggression and cyberbullying*, pages 32–41.

Samrat Thapa, Ashish Shah, Fawaz Ahmed Jafri, and et al. 2022. A multi-modal dataset for hate speech detection on social media: Case-study of russia-ukraine conflict. In *CASE 2022-5th Workshop on Challenges and Applications of Automated Extraction of Socio-Political Events from Text, Proceedings of the Workshop*. Association for Computational Linguistics.

Michael Wiegand, Melanie Siegel, and Roman Bock. 2019. Automated hate speech detection in multimodal data. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2973–2983.

Jiawei Xu, Yuxin Wang, Qi Liu, et al. 2023. Hypergraph neural networks for multimodal information fusion. *IEEE Transactions on Neural Networks and Learning Systems*, 34(6):2587–2600.

Chenguang Yang, Feida Zhu, Guoqing Liu, and et al. 2022. Multimodal hate speech detection via cross-domain knowledge transfer. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 4505–4514.

Linhao Zhang, Li Jin, Xian Sun, Guangluan Xu, Zequn Zhang, Xiaoyu Li, Nayu Liu, Qing Liu, and Shiyao Yan. 2023. Tot: topology-aware optimal transport for multimodal hate detection. In *Thirty-Seventh AAAI Conference on Artificial Intelligence, AAAI 2023, Thirty-Fifth Conference on Innovative Applications of Artificial Intelligence, IAAI 2023, Thirteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2023, Washington, DC, USA, February 7-14, 2023*, pages 4884–4892. AAAI Press.

Yuhao Zhou, Zeyu Chen, and Hang Yang. 2021. Multimodal learning for hateful memes detection. In *2021 IEEE International Conference on Multimedia & Expo Workshops (ICMEW)*, pages 1–6. IEEE.