Just *KIDDIN'* : *K*nowledge *I*nfusion and *D*istillation for *D*etection of *IN*decent Memes

Warning: This paper contains content that may be potentially offensive or upsetting.

Anonymous ACL submission

Abstract

Toxicity identification in online multimodal environments remains a challenging task due to the complexity of contextual connections across modalities (e.g., textual and visual). In this paper, we propose a novel framework that integrates Knowledge Distillation (KD) from Large Visual Language Models (LVLMs) and knowledge infusion to enhance the performance of toxicity detection in hateful memes. Our approach extracts sub-knowledge graphs from ConceptNet, a large-scale commonsense Knowledge Graph (KG) to be infused within a compact VLM framework. The relational context between toxic phrases in captions and memes, as well as visual concepts in memes enhance the model's reasoning capabilities. Experimental results from our study on two hate speech benchmark datasets demonstrate superior performance over the state-of-the-art baselines across AU-ROC, F1, and Recall with improvements of 1.1%, 7%, and 35%, respectively. Given the contextual complexity of the toxicity detection task, our approach showcases the significance of learning from both explicit (i.e. KG) as well as implicit (i.e. LVLMs) contextual cues incorporated through a hybrid neurosymbolic approach. This is crucial for real-world applications where accurate and scalable recognition of toxic content is critical for creating safer online environments. Our code is available at Anonymous Github[1].

1 Introduction

The rapid expansion of online platforms has led to an unprecedented increase in harmful and toxic content, posing significant challenges to maintaining safe and inclusive digital environments. According to the PEW Research Center, 41% of Americans have personally experienced online harassment (Vogels, 2021), while a recent study reported a surge in hate speech targeting religion during the COVID-19 pandemic (Chandra et al., 2021). Online toxic content, particularly in multimodal forms like memes, is often nuanced and context-dependent, making its detection a complex task (Kiela et al., 2020b; Sheth et al., 2022).

The contextual nature of toxicity poses a major challenge for detection. Pavlopoulos et al. (2020) found that 10.6% of posts labeled toxic in isolation were later judged non-toxic when viewed in full conversation context. Harmless-looking memes can also convey harmful messages through sarcasm, irony, or cultural references, potentially reinforcing stereotypes or negative societal norms (Waseem and Hovy, 2018; Gonzalez and Smith, 2020). Accurate toxicity assessment requires understanding the complete context, as implicit evaluations can misinterpret tone or intent, underscoring the importance of using explicit context in toxicity detection tasks (Sheth et al., 2022).

While the problem of online multimodal toxicity detection has witnessed advancements, several limitations remain that our approach aims to address. Existing approaches, including HateCLIPper (Kumar and Nandakumar, 2022a) and PromptHate (Cao et al., 2022), rely solely on training data and pre-trained models without incorporating external knowledge sources, limiting their ability to capture complex contextual cues and nuances of toxicity. On the other hand, large multimodal models, such as Flamingo (Alayrac et al., 2022) and LENS (Berrios et al., 2023), achieve comparable performance but are computationally expensive and challenging to deploy in resource-constrained environments. Existing models often struggle with nuanced or context-dependent toxicity because they rely on pattern matching rather than broader contextual knowledge (Pavlopoulos et al., 2022).

Our method addresses these limitations by leveraging an external commonsense KG, ConceptNet (Speer et al., 2017), thereby enhancing contextual comprehension and detecting context-dependent

[1] https://anonymous.4open.science/r/MultimodalToxicity-124F/README.md

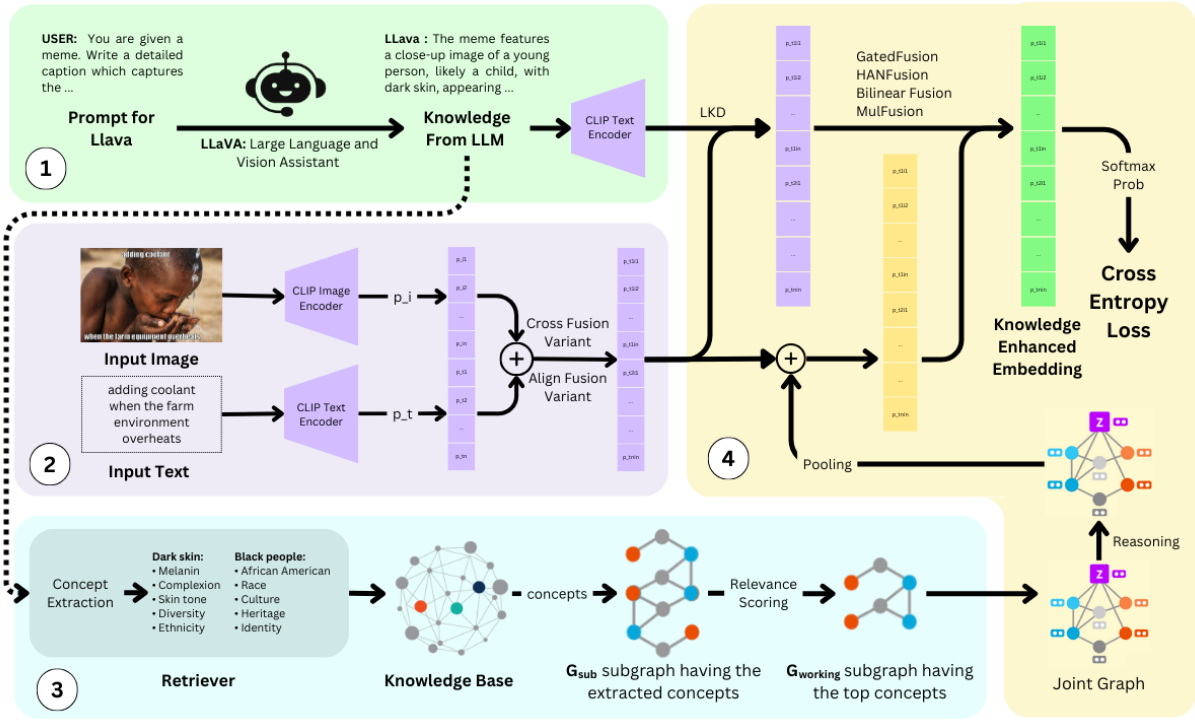


Figure 1: KID-VLM framework: The framework combines KD from a LVLM with knowledge infusion from external KGs such as ConceptNet. The input image and text are processed through the CLIP encoders to generate embeddings, which are fused using different fusion mechanisms. ① Knowledge Extraction from Teacher Model. ② Multimodal Learning Framework. ③ Knowledge Extraction from KG. ④ Joint Reasoning Space: reasoning using the implicit knowledge from teacher model and explicit knowledge from KG for toxicity prediction.

toxic content.

In this study, we propose the following research questions to guide our investigation into improving multimodal toxicity detection:

RQ1: How can we design a framework that distills knowledge from LVLMs as well as from Commonsense KG such as a ConceptNet to enhance smaller models' reasoning and improve their ability to detect toxicity in memes by combining both implicit and explicit contextual cues?

RQ2: How does the proposed knowledge-infused and distilled framework perform compared to existing multimodal toxicity detection models on benchmark datasets: *Hateful Meme* (Kiela et al., 2020b) and on*HarMeme* Dataset for generalizability (Pramanick et al., 2021a)?

We propose a novel approach that combines KD from LVLMs with knowledge extraction from KGs, integrating implicit and explicit contextual knowledge for more efficient *compact* VLMs. Using CLIP (Radford et al., 2021) as the backbone, we extract visual and textual features from memes, with one text encoder for captions generated by the LLaVA teacher model (Liu et al., 2024) and another for meme text. The visual and meme-text embeddings are fused using Align Fusion (Kumar and Nandakumar, 2022b), computing a Feature Interaction Matrix (FIM) to align visual and textual data. We refine the aligned multimodal representation by distilling knowledge from the teacher model using consistency loss. The student model learns to capture implicit contextual cues by aligning its internal representations with those of the teacher. We further incorporate external knowledge from ConceptNet via graph-based reasoning, constructing a joint graph that combines the meme context and KG entities. The fused multimodal and graph-based representations are then used for the final toxicity prediction.

Our findings, evaluated on two benchmark datasets, demonstrate the efficacy of this approach. The *Knowledge-Infused Distilled Vision-Language Models (KID-VLM)* framework consistently outperforms the baselines as the framework's dual-context learning approach (see Figure 1) enables the model to capture both the latent patterns from LVLMs and the explicit relational knowledge from KGs, resulting in a more comprehensive under-

standing of multimodal toxic content. Error analysis and ablation studies further underscore the importance of these components, highlighting the nuanced improvements in accuracy and contextual understanding.

2 Related Work

2.1 Online Toxicity Detection

Early approaches to toxicity detection primarily focused on textual content, employing traditional machine learning (Leo et al., 2023; Saha et al., 2023) and, later, advanced deep learning techniques were introduced (Jonathan and Setiawan, 2023; Karim et al., 2022) as per the nuanced and context-dependent nature of the online toxicity problem. Multimodal approaches have emerged in recent years that leverage both textual and visual content. Specifically, HateCLIPper (Kumar and Nandakumar, 2022a) and PromptHate (Cao et al., 2022) have shown promising results by utilizing pre-trained models, fusion methods, prompt tuning, and large datasets. Researchers introduced another multimodal framework for detecting toxicity in code-mixed videos through cross-modal synchronization (Maity et al., 2024). Large multimodal models such as Flamingo (Alayrac et al., 2022) and LENS (Berrios et al., 2023) have further improved the detection performance while their computational complexity poses challenges for deployment in resource-constrained environments, limiting their practical applicability.

2.2 Knowledge Infusion and Distillation

Knowledge Infusion (KI) is a paradigm that integrates external knowledge sources (e.g., KGs) into machine learning models to enhance the representation and improve reasoning abilities by capturing complex relationships in multimodal content, such as images, text, and videos. Incorporating structured knowledge provides models with access to the explicit context that is often missing from the raw training data (Zhang et al., 2022). Liang et al. (2023) proposed dual Knowledge Distillation (KD) to enhance both multimodal monolingual and cross-lingual summarization while filtering irrelevant visual features. Dai et al. (2022) introduced vision-language KD (VLKD) to improve dual-stream models for multimodal generation, boosting zero-shot performance while preserving text generation abilities. In health informatics, Sharma et al. (2019) integrated domain-specific knowledge to enhance the performance of predictive models for health-related tasks. Similarly, Mitra and et al. (2019) illustrated the utility of external knowledge in dialogue systems, showing improved contextual understanding. Xu et al. (2024) used KGs and LLMs to enhance synthetic clinical data generation, improving performance while addressing privacy and fairness concerns. (Lymperaiou et al., 2022) reviewed knowledge-enhanced multimodal learning, highlighting the role of KGs in improving commonsense, and temporal reasoning in VL models.

Knowledge distillation (KD) is a widely used technique where a smaller student model mimics a larger teacher model to improve efficiency while retaining performance. In multimodal learning, Wang et al. (2020) demonstrated KD's effectiveness in handling incomplete modalities by distilling knowledge from multimodal models. KD has also been applied in graph-based tasks, such as Hong et al. (2023), which improved GNN performance by capturing community structures.

In contrast to prior work, our approach combines KD with KI by leveraging KGs, as recent past work on online multimodal toxicity detection is limited by their reliance on training data without incorporating external knowledge sources, potentially constraining their ability to capture complex, context-dependent toxicity.

3 Methodology

Our approach, KID-VLM as illustrated in Figure 1, starts by generating multimodal feature representations, which are learned from a large pre-trained teacher model (e.g., LLaVA). These representations are then enriched by integrating knowledge from ConceptNet, resulting in a knowledge-enhanced representation. The subsequent sections outline the key processes of KD and joint reasoning in greater details.

3.1 Multimodal Representation: Aligned & Distilled from LVLM

We use CLIP as the backbone to extract visual and textual features from meme data. The vision encoder processes meme images, while two separate text encoders handle captions and OCR text. The *Caption Text Encoder* processes captions generated by the teacher model (LLaVA), with its weights frozen during training. For each image-text pair $D = \{I_i, T_i\}$, captions C_i are generated using the *LLaVA* model to capture the implicit meme context

(prompts are provided in Appendix Table 8). The *Caption Text Encoder*, *Meme Text Encoder*, and *Vision Encoder* generate embeddings for C_i, T_i, and I_i. Embeddings T_i, and I_i are then fused using the *Align Fusion* method (Kumar and Nandakumar, 2022b), which computes a *Feature Interaction Matrix (FIM)*, modeling interactions between image and text feature spaces:

$$R = \mathbf{p}_i \otimes \mathbf{p}_t \tag{1}$$

Diagonal elements of R represent the alignment between image and text embeddings:

$$\mathbf{p}_i^\top \mathbf{p}_t = \sum_{j=1}^{n} R_{jj} \tag{2}$$

This aligned representation, $\mathbf{w}_i$, is refined through KD, using a consistency loss defined as the Euclidean distance between the aligned embedding $\mathbf{w}_i$ and the embedding from the teacher model $\mathbf{w}_i^{clip_llava}$:

$$\mathcal{L}_{KD} = \|\mathbf{w}_i - \mathbf{w}_i^{clip_llava}\|_2^2 \tag{3}$$

This aligns the internal feature representations between the student and teacher models, enabling the student model to capture rich semantic knowledge and implicit contextual cues from the larger teacher model.

3.2 Graph-Based Reasoning

To enrich the distilled multimodal representation with explicit relational knowledge, we use a joint reasoning framework, which extracts working graphs for each data point using concepts from the meme caption and the LVLM-generated caption C_i, and encodes through graph neural networks.

3.2.1 Construction of the Working Graph

We first extract a subgraph G_{sub} (refer Figure 1) from ConceptNet based on the concepts from the meme caption and the LVLM caption. The nodes in this subgraph are ranked by relevance scoring ρ_v to reduce noise and focus on the most relevant KG entities. ρ_v for each node $v \in V_{\text{sub}}$ (nodes of G_{sub}) is calculated based on its relationship to the meme context. We create a context node z encapsulating the overall context of the data point. The score is calculated by passing the concatenated text representations of z and v through the Roberta (Liu et al., 2019) model, where the relevance score is determined using the perplexity score:

$$\rho_v = f_{\text{head}}(f_{\text{enc}}([\text{text}(z); \text{text}(v)])) \tag{4}$$

Additionally, for experimentation, we compute the cosine similarity between the concatenated text representation of z and KG entities v using MiniLM (Wang et al., 2021) (Sentence-Transformer) to obtain the Relevancy Score $\rho_v = \text{similarity}(z, v)$.

We use these two scores separately for the experiments to assess the relevance, retaining the top k (i.e., 750) most relevant entities for inclusion in the working graph G_{W} (G_{working} in Figure 1). After constructing the initial working graph, we add the context node z to integrate the two sources of knowledge—from the LVLM and KG. This context node is then connected to the entities V_{sub} from the working graph with new relationship r', forming the final working graph G_{W}, to prepare for knowledge representation learning in the next step.

3.2.2 Working Graph Knowledge Representation

We experiment with two different graph neural network architectures for reasoning over the joint working graph G_{W}: the *Relational Graph Convolutional Network (R-GCN)* (Schlichtkrull et al., 2017) and the *Graph Attention Network (GAT)* (Veličković et al., 2018). Each architecture is evaluated independently to explore its effectiveness in integrating multimodal representations and KG relationships. While R-GCN applies relation-specific transformations to process messages, GAT uses attention mechanisms to weigh them. In R-GCN, node representations are updated by aggregating information from neighboring nodes based on relation-specific transformations:

$$h_v^{(l+1)} = \sigma\left(\sum_{r \in R} \sum_{u \in \mathcal{N}_r(v)} W_r^{(l)} h_u^{(l)}\right) \tag{5}$$

where $W_r^{(l)}$ is the relation-specific transformation matrix and $\mathcal{N}_r(v)$ are neighbors connected by relation r. In GAT, attention coefficients α_{vu} are computed between a node v and its neighbor u, allowing the model to focus on the most important neighbors:

$$\alpha_{vu} = \text{softmax}\left(\text{LeakyReLU}\left(\mathbf{a}^\top [Wh_v \parallel Wh_u]\right)\right) \tag{6}$$

The node representation is updated by aggregating the weighted sum of its neighbors.

Graph Pooling: Once the node representations h_v for all nodes $v \in G_{\text{W}}$ are updated through graph encoding (e.g., R-GCN, GAT), we learn a pooled

Framework	Knowledge	Accuracy	F1	Precision	Recall	AUC
MMBT	-	68.80	67.61	71.25	68.44	77.84
MMBT	Triples	69.40	68.04	72.46	69.02	78.28
MMBT	Summarized Triples	69.60	68.96	70.83	69.34	78.03
CLIP-Text*	-	55.30	46.63	70.09	15.31	65.75
CLIP-Image*	-	62.20	58.98	74.35	34.89	73.42
CLIP*	-	59.50	53.88	76.40	25.10	72.14
CLIP-Text	-	59.90	56.87	68.16	34.08	68.01
CLIP-Image	-	55.70	53.55	57.97	34.90	60.78
CLIP	-	60.30	57.10	69.62	33.67	68.30
LLaVA+	-	57.00	52.43	60.00	56.40	-
HateClipper	-	76.07	73.00	78.00	56.00	85.14
KID-VLM	Hop 1 & Roberta	**78.20** (+2.13)	**78.12** (+5.12)	80.36(+2.36)	73.47 (+17.47)	86.02 (+0.88)
KID-VLM	Hop 2 & Roberta	76.30 (+0.23)	76.07 (+3.07)	76.94 (-1.06)	**76.14** (+20.14)	**86.11** (+0.97)
KID-VLM	Hop 1 & MiniLM	**78.70** (+2.63)	**78.63** (+5.63)	**80.57** (+2.57)	**74.49** (+18.49)	**86.11** (+0.97)
KID-VLM	Hop 2 & MiniLM	76.30(+0.23)	75.95 (+2.95)	**82.52** (+4.52)	65.51 (+9.51)	85.48 (+0.34)

Table 1: Performance Comparison of various models on Hateful Memes Dataset across multiple metrics. The KID-VLM framework shows superior results as compared to other model (**bold**). * represents Frozen CLIP encoders, + represents model was evaluated in zeroshot. The deltas are reported with respect to the HateClipper baseline.

representation of the graph to generate a single vector representing the entire working graph. We use a pooling operation that summarizes node-level embeddings into a global graph-level representation.

For graph pooling, we use Mean Pooling:

$$h_{\text{graph}} = \frac{1}{|V_{\text{W}}|} \sum_{v \in V_{\text{W}}} h_v \tag{7}$$

These pooling techniques condense the graph's information into a single vector h_{graph}, which is later combined with the distilled model's representation in the fusion mechanism.

Fusion Mechanism: The pooled graph representation h_{graph} is fused with the multimodal representation $h_{\text{distilled}}$ from the distilled model. We explored various fusion techniques (Appendix B,Table 4) to combine these representations. *Gated Fusion* (Xiong et al., 2024) emerged as the most effective, defined as:

$$\begin{aligned} \mathbf{G} &= \sigma\left(\mathbf{W}_g \left(h_{\text{graph}} \parallel h_{\text{distilled}}\right)\right) \\ \mathbf{F}_{\text{multimodal}} &= \mathbf{G} \odot h_{\text{graph}} + (1 - \mathbf{G}) \odot h_{\text{distilled}} \end{aligned} \tag{8}$$

where, h_{graph} is the pooled graph representation, $h_{\text{distilled}}$ is the multimodal embedding, and σ is the sigmoid activation function. The gating mechanism learns to balance the contributions of both representations.

Toxicity Prediction: Final toxicity prediction is made using the fused representation $\mathbf{F}_{\text{multimodal}}$ from the graph and multimodal embeddings. The probability of toxicity is given by:

$$p_{\text{tox}} = \sigma\left(W_{\text{tox}}\left[\mathbf{F}_{\text{multimodal}}\right] + b_{\text{tox}}\right) \tag{9}$$

where $\mathbf{F}_{\text{multimodal}}$ is the fused representation, W_{tox} is the weight matrix, b_{tox} is the bias term, and σ is the sigmoid activation function.

Loss Function: The training process optimizes a joint loss function with two components: Binary Cross-Entropy Loss (BCE) for meme classification and Consistency Loss to enforce consistency between the multimodal fusion and captions from the teacher model. The total loss is a weighted combination of these two losses:

$$\mathcal{L}_{\text{total}} = \lambda_1 \mathcal{L}_{\text{BCE}} + \lambda_2 \mathcal{L}_{\text{KD}} \tag{10}$$

which controls the balance between the two losses. The losses are weighted using hyperparameters, and the overall loss is minimized using the AdamW optimizer.

3.3 Experiments

We evaluate the performance of *KID-VLM* on two benchmark datasets primarily used for Hatespeech Detection on the HatefulMemes dataset (Kiela et al., 2020b) and the HarMeme dataset (Pramanick et al., 2021b) (Dataset statistics are provided in Appendix A) For fair comparison, we adopt the evaluation metrics used in prior works (Kumar and Nandakumar, 2022a; Cao et al., 2022), including AUC, Recall and Accuracy. In our experiments with the KID-VLM framework, several key design decisions were made to effectively utilize

Framework	Target Metrics				Intensity Metrics			
	Acc.	MA-F1	Precision	Recall	Acc.	MA-F1	Precision	Recall
CLIP-Text	67.74	54.39	41.33	37.61	75.14	50.09	48.45	53.38
CLIP-Image	62.10	36.41	60.57	39.24	80.23	53.54	51.48	56.89
CLIP	57.26	35.67	50.29	39.24	80.23	53.60	51.55	57.06
CLIP-Text*	59.68	35.43	74.42	52.17	68.36	45.13	43.70	47.94
CLIP-Image*	62.10	36.41	60.57	39.24	64.97	26.26	21.66	33.33
CLIP*	62.10	36.88	39.86	38.88	77.12	50.59	49.03	52.26
HateClipper	75	73.42	75.83	**79.66**	79.66	78.76	75.83	79.66
KID-VLM (Hop 2)	**77.42**(+2.42)	**75.88**(+2.46)	**78.51**(+2.68)	77.42(-2.22)	**81.07** (+1.41)	**79.08** (+0.32)	77.73(+1.90)	**81.07** (+1.41)
KID-VLM (Hop 1)	**77.42**(+2.42)	**75.88**(+2.46)	**78.51** (+2.68)	77.42(-2.22)	80.51 (+0.85)	78.73 (-0.03)	**78.10** (+2.27)	80.51 (+0.85)

Table 2: Performance Comparison of models on HarMeme Dataset across multiple metrics. The KID-VLM framework, using LLaVA for caption generation and MiniLM for relevancy scoring, outperforms other models (**bold**). * represents Frozen CLIP encoders. The deltas are reported with respect to the HateClipper baseline.

KGs and fusion mechanisms. ConceptNet was employed as an external source of explicit contextual information, and we leveraged multi-hop pathway traversal to expand the model's understanding of relationships between concepts. Specifically, we experimented with *Hop 1* and *Hop 2* expansions. Additionally, we explored different values of top_k nodes, where k ranged from 250 to 750, to select the most relevant nodes for inclusion. For fusion of the multimodal representations with the graph-based representation, we evaluated multiple fusion mechanisms, including *Gated Fusion*, *Multiplicative Fusion*, *Bilinear Fusion*, and *Hierarchical Attention Fusion* (details in Appendix B). For processing the G_w, we employed two types of graph networks: *GCN* and *GAT*.

3.4 Baselines

We utilize multiple baseline *compact* VLMs on the two benchmark datasets to compare the performance of our approach. We included HateClipper (Kumar and Nandakumar, 2022b), an established benchmark derived from the Hate Memes Challenge[2], the Multimodal Bitransformer (MMBT) model (Kiela et al., 2020a), CLIP-based models (Radford et al., 2021), specifically CLIP-Text and CLIP-Image. Further, we created versions of the CLIP model, labeled as CLIP*, where only the projection layers are fine-tuned. We also used the LLava in zero-shot settings for a comparison with an LVLM.

3.5 Implementation Details

We use Optuna for hyperparameter tuning. We sweep through GNN hidden dimensions, number of layers, learning rate, fusion methods, and dropout rates. The model is trained for 30 epochs using the AdamW optimizer with learning rate warm-up and linear decay. A batch size of 4 is used on a RTX 5000 with VRAM size of 32 GBs, and model performance is monitored on the validation set during training, selecting the best model based on validation AUC. The GNN hidden and output dimensions are set within ranges of 2 to 512 and 2 to 1024, respectively. The mapping layer dimension is set between 2 and 2048, the learning rate ranges from 1e-10 to 1e-2, and dropout probabilities are selected within the range of 0.0 to 0.9. Additionally, we tune the number of mapping and pre-output layers between 1 and 5, weight decay between 1e-8 and 1e-1, and the loss alpha between 0 and 1.

4 Results & Discussion

Results on the *HateMeme* and *HarMeme* datasets (see Tables 1 and 2) show that by integrating external knowledge from ConceptNet and distilling multimodal information through the LLaVA model, *KID-VLM* framework consistently surpasses baseline models *HateCLIPper* and standard *CLIP* architectures across AUC, F1, and Recall with highest improvements of 1.1%, 7%, and 35%, respectively for *Hateful Memes* dataset.(Dataset details in Appendix A). A similar improvement was observed on the *HarMeme* dataset, with approximately 3% gains in Accuracy, MA-F1, and Precision.

4.1 Performance on HateMeme Dataset

On the *HateMeme* dataset, *KID-VLM*, with *Hop 1* sub-graphs and Gated Fusion achieved the highest AUC of *86.11* with accuracy of *78.20%*, and F1 score of *78.12*, outperforming *LLaVA* and *HateCLIPper*. In addition, *KID-VLM* with *Hop 2* achieves the highest recall of *76.14%* while maintaining the highest AUC of *86.11*, the accuracy of *76.30%*, highlighting the effectiveness of knowl-

[2] https://paperswithcode.com/sota/meme-classification-on-hateful-memes

edge infusion and distillation. Relevancy scoring using *MiniLM*(Sentence Transformer, another approach used for relevancy scoring apart from Roberta) for the creation of working graph G_w from ConceptNet resulted in a marginal decrease in Hop 1 accuracy (*73.80%*), but improved precision, recall, and AUC, suggesting enhanced generalization. The combination of distilled multimodal representations and external KGs evidently contributes to this overall performance improvement.

4.2 Performance on HarMeme Dataset

On the *HarMeme* dataset, *KID-VLM* Hop 2 achieved the best performance with an F1 score of *79.08%* and recall of *81.07%*, surpassing the baseline *HateClipper* model. By leveraging ConceptNet subgraphs and the KD process, *KID-VLM* effectively captures the nuances of hateful content, even across varying levels of intensity. Gains in recall, particularly in detecting harmful content, demonstrate the value of external knowledge in meme classification.

4.3 Better Recall using *KID-VLM*

Recall plays a critical role in toxicity detection, particularly when the model cannot detect actual toxic content as it may have potentially grave consequences. In the *Hate Meme* dataset, the *KID-VLM Hop 2* model achieved a recall score of *76.14%*, 35% and 10% higher than HateClipper and MMBT, respectively. This improvement is significant concerning real-world scenarios where false negatives (i.e., failing to detect toxic content) can have adverse impacts on individuals being exposed to potentially toxic content. If such a model were deployed by an online platform, this impact could be amplified towards a bigger audience. In this context, a high recall indicates that the *KID-VLM* model is more adept in identifying the subtle forms of toxicity that might be missed by other models. This is essential for detecting harmful content in nuanced contexts, such as memes that rely on cultural references, sarcasm, or visual metaphors. By leveraging external knowledge from ConceptNet and the KD from the teacher model, the model benefits from a broader contextual understanding, which enhances its ability to generalize and detect toxicity across different datasets.

4.4 Impact of Node Count on Performance

As seen in Table 3, increasing the number of nodes from KG has a noticeable impact on the AUC. When the node count is set to 250, the model achieves an AUC of 84.91, while it improves to 85.76 with 500 nodes. Node count 750 yields the highest AUC of 86.11. Increasing the amount of knowledge extracted from the ConceptNet enables the model to better capture the underlying contextual relationships between concepts in the meme, thereby improving its reasoning capability and ultimately boosting its performance. While the inclusion of additional nodes may likely provide additional context, the relevance needs to be maintained to avoid noise.

Framework	# Nodes	AUC
KID-VLM	250	84.91
KID-VLM	500	85.76
KID-VLM	750	**86.11**

Table 3: Ablation study on the impact of node count in hop-1 ConceptNet subgraphs on HateMEME dataset. Models use an RGCN encoder and Gated Fusion with CLIP as visual/text encoder. Performance of KID-VLM improves with increasing node count

Framework	Fusion	AUC
KID-VLM	Gated Fusion	**86.11**
KID-VLM	Multiplicative Fusion	80.64
KID-VLM	Bilinear Fusion	85.26
KID-VLM	Hierarchical Attention	85.39

Table 4: Ablation study of fusion mechanisms (Gated, Multiplicative, Bilinear, Hierarchical Attention) for KID-VLM on HateMeme Dataset using hop-2 ConceptNet (750 nodes) and CLIP encoders. Gated Fusion Mechanism shows a much better performance as compared to others on AUC (bold).

Fusion Mechanisms & GNN Architecture. The use of *Gated Fusion* within *KID-VLM* (Table 4) demonstrates the model's ability to integrate multimodal features effectively. Results show that the gating mechanism allows the model to balance between visual and textual information, boosting performance across both datasets. Moreover, we experiment with two GNN architectures, and *KIDVLM* with GCN produced the highest AUC scores of **86.11**.

4.5 Interpretability of Knowledge-Enhanced Models:

One of the key advantages of incorporating KGs into our toxicity detection framework is the enhanced interpretability of the model's predictions. By leveraging ConceptNet, the model is able to reason more effectively about the relationships both within and between textual and visual elements present in the memes. The nodes extracted from ConceptNet provide semantic context for under-

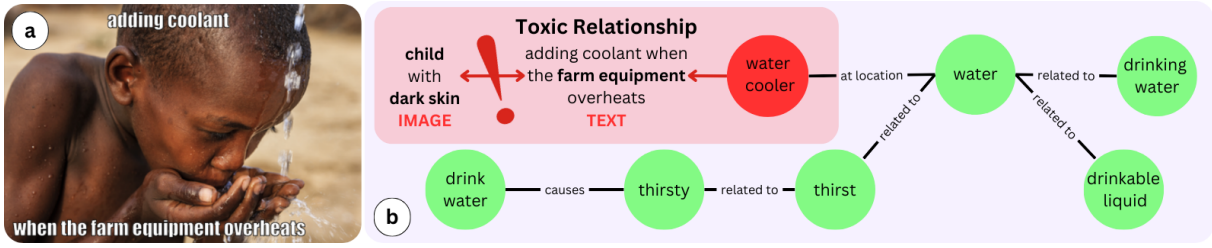


Figure 2: ⓐ **Generated LLaVA Caption:** The meme features a close-up image of a young person, likely a ***child, with dark skin***, appearing to be a boy. He is captured in the midst of drinking water from his hands, with water droplets splashing around his face. The person is looking down, seemingly focused on the water. The meme is captioned with the text ***adding coolant when the farm equipment overheats***. This caption is a humorous take on the concept of ***adding coolant*** to a situation that is not related to machinery or vehicles.
ⓑ **Subgraph** from ConceptNet showing the semantic relationships between terms related to the meme's content. Key terms such as ***"drinking water", "coolant" and "thirst"*** are connected to other concepts in the graph. For example, ***"drink water"*** is linked to ***"thirst" and "water cooler,"*** representing how these entities are conceptually related. This subgraph helps the model understand both the literal and metaphorical connections within the meme, allowing it to detect potentially toxic associations.

standing both literal and metaphorical meanings in the memes. For example, as illustrated in Figure 2, the KG helps the model establish connections between key terms such as *"drinking water," "coolant," and "thirst"*. These concepts are not only linked to their literal meanings but also provide insights into the metaphorical associations present in the meme, such as *"overheating"* being compared to feeling overwhelmed. By grounding the model's understanding in such knowledge-driven relationships, the reasoning process becomes more transparent. Thus, the inclusion of KGs not only boosts model performance but also significantly enhances the interpretability of the model's predictions. The ability to visualize and understand the reasoning behind toxicity classification is a major step toward building more explainable AI systems for multimodal content analysis (for more details read Appendix C).

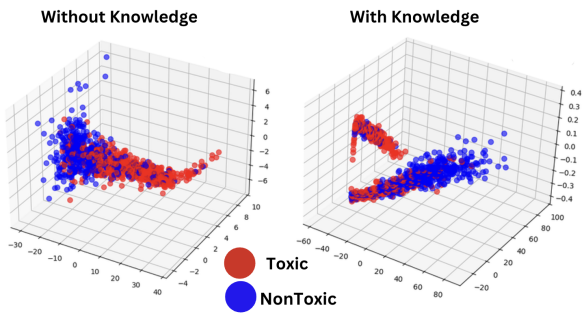


Figure 3: Baseline vs. KID-VLM t-SNE plots illustrating the reduced 3D representation of the dataset after dimensionality reduction. The colors represent the ground truth labels of the data points. KID-VLM's plot shows a much clearer separation between the labeled data points.

4.6 Clearer Separation: Toxic vs. Non-Toxic

Enhancing representations by knowledge infusion leads to a more distinct separation between toxic and non-toxic content. Figure 3 shows how multimodal data points are placed with (right) and without enhancement (left), demonstrating the difference in separation in the latent space. Without the knowledge-enhanced representations, the boundary between these two classes is more ambiguous and the data points are more scattered, whereas the knowledge-enhanced representations were separated into distinct regions of the latent space with denser clusters. As ambiguity and unclear separation often lead to misclassifications, especially in edge cases where the context is crucial for understanding whether a meme contains harmful content, the knowledge-infused approach provides improved robustness. This clear separation may stem from both explicit and implicit contextual cues learned during the training.

5 Conclusion

In summary, the results demonstrate that the *KID-VLM* architecture, through its combination of KGs and KD, provides superior performance for detecting hate speech in memes compared to baseline methods. The use of *ConceptNet subgraphs* and distilled multimodal representations allows the model to better understand the complex semantics of multimodal content, resulting in improved accuracy, F1, and AUC scores. These findings highlight the critical role of knowledge infusion and advanced fusion mechanisms in enhancing multimodal hate speech detection systems.

6 Limitations & Future Work

While *KID-VLM* demonstrates strong performance on the *HateMeme* and *HarMeme* datasets, it has several limitations. The model's reliance on ConceptNet for external knowledge may limit its effectiveness in cases where the graph lacks relevant or niche information, and its generalizability to other datasets beyond these two has not been extensively tested. Additionally, incorporating graph-based methods increases computational complexity, which may hinder scalability to larger datasets or real-time applications. The quality of KD from larger models could degrade when using smaller student models, and there remains a potential risk of bias from pretrained models and KGs. Future work could address these concerns by exploring more diverse datasets, improving scalability, and investigating bias mitigation strategies.

7 Ethical Considerations

The goal of this study is to improve toxicity detection in multimodal environments by combining KD from Large Visual Language Models (LVLMs) and knowledge infusion from commonsense KGs. However, toxicity detection systems can mislabel content or users, especially when dealing with complex inputs like memes that involve irony or satire. Cultural and social context may not be fully understood by the model, so continuous refinement is essential to reduce biases. The study uses publicly available anonymized datasets, but potential misuse of such systems in surveillance or censorship remains a concern. Responsible usage, with clear guidelines to protect free expression, is critical. Additionally, the potential for LVLMs to propagate biases from their training data must be addressed through ongoing evaluation and the use of diverse data sources.

References

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katie Millican, Malcolm Reynolds, et al. 2022. Flamingo: a visual language model for few-shot learning. *Advances in Neural Information Processing Systems*, 35:23716–23736.

William Berrios, Yonglong Tian, Chiraag Lala, Alec Parrish, and Trevor Darrell. 2023. Lens: A learnable embedding name space for multi-modal, multi-task learning. *arXiv preprint arXiv:2306.11527*.

Ruijun Cao, Jing Chen, Xin Guo, Chaoyu Li, Yanan Ma, Zhongyu Liu, Zhichao Cheng, and Yang Gao. 2022. Prompthate: A knowledge-aware prompt-based fine-tuning framework for hate speech detection. *arXiv preprint arXiv:2207.14344*.

Mohit Chandra, Manvith Reddy, Shradha Sehgal, Saurabh Gupta, Arun Balaji Buduru, and Ponnurangam Kumaraguru. 2021. "a virus has no religion": Analyzing islamophobia on twitter during the covid-19 outbreak. In *Proceedings of the 32nd ACM Conference on Hypertext and Social Media*, HT '21, page 67–77, New York, NY, USA. Association for Computing Machinery.

Wenliang Dai, Lu Hou, Lifeng Shang, Xin Jiang, Qun Liu, and Pascale Fung. 2022. Enabling multimodal generation on CLIP via vision-language knowledge distillation. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2383–2395, Dublin, Ireland. Association for Computational Linguistics.

Maria Gonzalez and John Smith. 2020. The role of memes in online discourse: A study on toxicity and humor. *Journal of Digital Communication*.

Chao Hong, Huanhuan Yin, Weijia Zhang, and Xiangliang Zhang. 2023. Research on community detection method based on knowledge distillation. *IEEE Transactions on Network Science and Engineering*.

Vincent Williams Jonathan and Erwin Budi Setiawan. 2023. Feature Expansion Using GloVe for Hate Speech Detection using Convolutional Neural Network (CNN) and Recurrent Neural Network (RNN) Method in Twitter. *2023 International Conference on Data Science and Its Applications (ICoDSA)*.

Md Rezaul Karim, Sumon Kanti Dey, Tanhim Islam, Md Shajalal, and Bharathi Raja Chakravarthi. 2022. Multimodal Hate Speech Detection from Bengali Memes and Texts. ArXiv:2204.10196 [cs].

Douwe Kiela, Suvrat Bhooshan, Hamed Firooz, Ethan Perez, and Davide Testuggine. 2020a. Supervised Multimodal Bitransformers for Classifying Images and Text. ArXiv:1909.02950.

Douwe Kiela, Hamed Firooz, Aravind Mohan, Vedanuj Goswami, Amanpreet Singh, Pratik Ringshia, and Davide Testuggine. 2020b. The hateful memes challenge: Detecting hate speech in multimodal memes. In *Advances in Neural Information Processing Systems*.

Gokul Karthik Kumar and Karthik Nandakumar. 2022a. Hate-clipper: Multimodal hateful meme classification based on cross-modal interaction of clip features. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 4823–4827.

Gokul Karthik Kumar and Karthik Nandakumar. 2022b. Hate-clipper: Multimodal hateful meme classification based on cross-modal interaction of clip features. *arXiv preprint arXiv:2210.05916*.

Chelsea Olivia Leo, Bagus Jati Santoso, and Baskoro Adi Pratomo. 2023. Enhancing Hate Speech Detection for Social Media Moderation: A Comparative Analysis of Machine Learning Algorithms. *2023 International Conference on Advanced Mechatronics, Intelligent Manufacture and Industrial Automation (ICAMIMIA)*, pages 960–964.

Yinhan Liang, Zhengyuan Zhang, Zichao Yang, et al. 2023. D2tv: Document-to-tv-commercial video summarization via knowledge-infused multimodal learning. *arXiv preprint arXiv:2301.00994*.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2024. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26296–26306.

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

Eirini Lymperaiou, Fabio Petroni, and Evangelos Kanoulas. 2022. A survey on knowledge-enhanced multimodal learning. *arXiv preprint arXiv:2211.12328*.

Krishanu Maity, A. S. Poornash, Sriparna Saha, and Pushpak Bhattacharyya. 2024. ToxVidLM: A Multimodal Framework for Toxicity Detection in Code-Mixed Videos. ArXiv:2405.20628.

Arindam Mitra and et al. 2019. Towards improving the performance of dialogue systems via knowledge-infused learning. *Proceedings of the 2019 AAAI Conference on Artificial Intelligence*, 33(01):7689–7696.

John Pavlopoulos, Leo Laugier, Alexandros Xenos, Jeffrey Sorensen, and Ion Androutsopoulos. 2022. From the detection of toxic spans in online discussions to the analysis of toxic-to-civil transfer. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3721–3734, Dublin, Ireland. Association for Computational Linguistics.

John Pavlopoulos, Jeffrey Sorensen, Lucas Dixon, Nithum Thain, and Ion Androutsopoulos. 2020. Toxicity detection: Does context really matter? *arXiv preprint arXiv:2006.00998*.

Shraman Pramanick, Dimitar Dimitrov, Rituparna Mukherjee, Shivam Sharma, Md Shad Akhtar, Preslav Nakov, and Tanmoy Chakraborty. 2021a. Detecting harmful memes and their targets. *arXiv preprint arXiv:2110.00413*.

Shraman Pramanick, Dimitar Dimitrov, Rituparna Mukherjee, Shivam Sharma, Md. Shad Akhtar, Preslav Nakov, and Tanmoy Chakraborty. 2021b. Detecting harmful memes and their targets. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 2783–2796, Online. Association for Computational Linguistics.

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

Sougata Saha, Michael Sullivan, and R. Srihari. 2023. Hate Speech Detection in Low Resource Indo-Aryan Languages.

Michael Schlichtkrull, Thomas N. Kipf, Peter Bloem, Rianne van den Berg, Ivan Titov, and Max Welling. 2017. Modeling Relational Data with Graph Convolutional Networks. ArXiv:1703.06103.

Priyanshu Sharma, Vivek Chalasani, Tanmay Agrawal, Matanel Mor, and Shubham Somani. 2019. Incorporating domain knowledge into deep learning models for clinical predictions. In *Proceedings of the 2019 IEEE International Conference on Healthcare Informatics (ICHI)*, pages 1–7. IEEE.

Amit Sheth, Valerie L Shalin, and Ugur Kursuncu. 2022. Defining and detecting toxicity on social media: context and knowledge are key. *Neurocomputing*, 490:312–318.

Robyn Speer, Joshua Chin, and Catherine Havasi. 2017. Conceptnet 5.5: An open multilingual graph of general knowledge. In *Thirty-First AAAI Conference on Artificial Intelligence*.

Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. 2018. Graph Attention Networks. ArXiv:1710.10903.

Emily A Vogels. 2021. The state of online harassment. *Pew Research Center*.

Qi Wang, Shafiq Joty, Steven Lyu, Irwin King, Caiming Xiong, and Steven CH Hoi. 2020. Multimodal learning with incomplete modalities by knowledge distillation. *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 1908–1918.

Wenhui Wang, Hangbo Bao, Shaohan Huang, Li Dong, and Furu Wei. 2021. MiniLMv2: Multi-head self-attention relation distillation for compressing pretrained transformers. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 2140–2151, Online. Association for Computational Linguistics.

Zeerak Waseem and Dirk Hovy. 2018. Hateful symbols or hateful people? predictive features for hate speech detection on twitter. In *Proceedings of the Second Workshop on Ethics in Natural Language Processing*.

Liyan Xiong, Xinhua Yuan, Zhuyi Hu, Xiaohui Huang, and Peng Huang. 2024. Gated fusion adaptive graph neural network for urban road traffic flow prediction. *Neural Processing Letters*, 56(1):9.

Ran Xu, Hejie Cui, Yue Yu, Xuan Kan, Wenqi Shi, Yuchen Zhuang, May Dongmei Wang, Wei Jin, Joyce Ho, and Carl Yang. 2024. Knowledge-infused prompting: Assessing and advancing clinical text data generation with large language models. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 15496–15523.

Xikun Zhang, Antoine Bosselut, Michihiro Yasunaga, Hongyu Ren, Percy Liang, Christopher D Manning, and Jure Leskovec. 2022. Greaselm: Graph reasoning enhanced language models for question answering. *International Conference on Representation Learning (ICLR)*.

A Datasets

A.1 Hateful Memes Dataset

The primary dataset used in our experiments is the Hateful Memes Challenge Dataset (Kiela et al., 2020b), consisting of 10K meme images accompanied by their respective text overlays. Each meme is labeled as either "Hateful" or "Non-hateful." The dataset is divided into training, testing, and validation subsets, as detailed in Table 5. Representative examples of memes from the dataset are presented in Figure 4.

Labels	**Train Set**	**Val Set**	**Test Set**
Non-Hateful (0)	5481	253	510
Hateful (1)	3019	247	490
Total	8500	500	1000

Table 5: Train, Test and Validation Seen splits for the Hateful Memes Dataset

A.2 HarMeme Dataset

We also conducted a series of experiments on the HarMeme Dataset (Pramanick et al., 2021b), a benchmark dataset for hateful meme classification. This dataset comprises 3,544 memes related to COVID-19, collected from the Internet, with each meme annotated with both [Intensity, Target] labels. The dataset is split into training, testing, and validation sets for the Intensity and Target labels, as detailed in Tables 6 and 7, respectively. Figure 5 presents several example memes from the dataset.

Labels	**Train Set**	**Val Set**	**Test Set**
Not Harmful (0)	1949	116	230
Somewhat Harmful (1)	882	51	103
Very Harmful (2)	182	10	21
Total	3013	177	354

Table 6: Train, Test and Validation splits for the HarMeme Dataset Intensity Variable

Labels	**Train Set**	**Val Set**	**Test Set**
Individual	493	30	59
Organisation	65	3	7
Community	279	16	32
Society	226	13	26
Total	1063	62	124

Table 7: Train, Test and Validation splits for the HarMeme Dataset Target Variable

B Fusion Mechanisms

Let $\mathbf{E}_g \in \mathbf{R}^{d_g}$ and $\mathbf{E}_m \in \mathbf{R}^{d_m}$ denote the graph pooled representation and the distilled multimodal representation, respectively, where d_g and d_m represent the dimensions of the graph and multimodal embeddings. Several fusion mechanisms are evaluated for combining these embeddings from different modalities:

Gated Fusion applies a gating mechanism to combine the graph and multimodal embeddings. The fusion is computed as:

$$\mathbf{F}_{\text{gated}} = \begin{array}{l} \sigma(\mathbf{W}_g[\mathbf{E}_g \| \mathbf{E}_m]) \odot \mathbf{E}_g \\ + (1 - \sigma(\mathbf{W}_g[\mathbf{E}_g \| \mathbf{E}_m])) \odot \mathbf{E}_m \end{array} \quad (11)$$

where $\mathbf{W}_g$ represents learnable weights, $\sigma(\cdot)$ is the sigmoid activation, and $\odot$ denotes element-wise multiplication.

Bilinear Pooling Fusion computes the bilinear interaction between the graph and multimodal embeddings:

$$\mathbf{F}_{\text{bilinear}} = \mathbf{E}_g^T \mathbf{W}_b \mathbf{E}_m \quad (12)$$

where $\mathbf{W}_b$ is a bilinear transformation matrix.

HAN Fusion (Hierarchical Attention Network) applies attention weights at multiple levels to the graph and multimodal embeddings:

$$\mathbf{F}_{\text{HAN}} = \sum_{l=1}^{L} \alpha_l (\mathbf{W}_l [\mathbf{E}_g \| \mathbf{E}_m]) \quad (13)$$

where α_l is the attention weight at level l and $\mathbf{W}_l$ is the learnable weight matrix.

Finally, **Multiplicative Fusion** combines the embeddings multiplicatively after non-linear transformation:

$$\mathbf{F}_{\text{mult}} = \tanh(\mathbf{W}_m \mathbf{E}_g) \odot \tanh(\mathbf{W}_m \mathbf{E}_m) \quad (14)$$

where $\tanh(\cdot)$ is the hyperbolic tangent activation, and $\mathbf{W}_m$ is a weight matrix.

C Interpretability Examples

By incorporating ConceptNet,the model's interpretability is enhanced, enabling it to reason more effectively about the relationships between textual and visual elements in memes. The extracted nodes from ConceptNet offer semantic context that aids in understanding both literal and metaphorical meanings, improving the model's ability to explain its reasoning.

As illustrated in Figure 6 , the LLaVA caption interprets the meme as contrasting the perceived ease of entering Islam, depicted as a joyful experience, with the challenges of leaving Islam, which are portrayed as distressing or dangerous. The top panel shows a celebratory embrace, while the bottom panel illustrates chaos and urgency surrounding an individual in distress. The knowledge sub-graph, containing nodes such as Islamophobia, Muslim, and Islam, underscores the religious context. When combined, the LLaVA explanation and the knowledge graph reveal how the meme perpetuates harmful stereotypes and fosters negative sentiments towards Islam, reinforcing its toxicity.

Similarly, in Figure 7, the LLaVA caption interprets the meme as addressing issues of racial discrimination and violence. The meme features a close-up image of a person with a bruised eye, implying a physical altercation. The individual, who has blonde hair, stares directly at the camera with a serious expression. The text on the image reads, "overheard my daughter telling her friend she was curious what it would be like to date a black guy so... I showed her!" The meme uses satire to comment on societal issues, specifically the harmful stereotypes and prejudices surrounding interracial relationships, particularly those involving Black individuals.The knowledge sub-graph includes nodes such as racial discrimination, racism, and domestic violence, underscoring the meme's connection to racial issues and violence. The meme critiques how prejudice and discrimination can manifest, especially when parents perpetuate such harmful views in response to their children's actions. While the meme employs dark humor and satire, it ultimately reflects on the damaging effects of racism and discrimination. It is important to note that the meme does not endorse violence but rather uses the image as a metaphor to highlight the harmful consequences of racial bias, addressing a broader societal issue.

The model's ability to visualize and comprehend the reasoning behind toxicity classification represents a significant advancement in developing more explainable AI systems for multimodal content analysis.

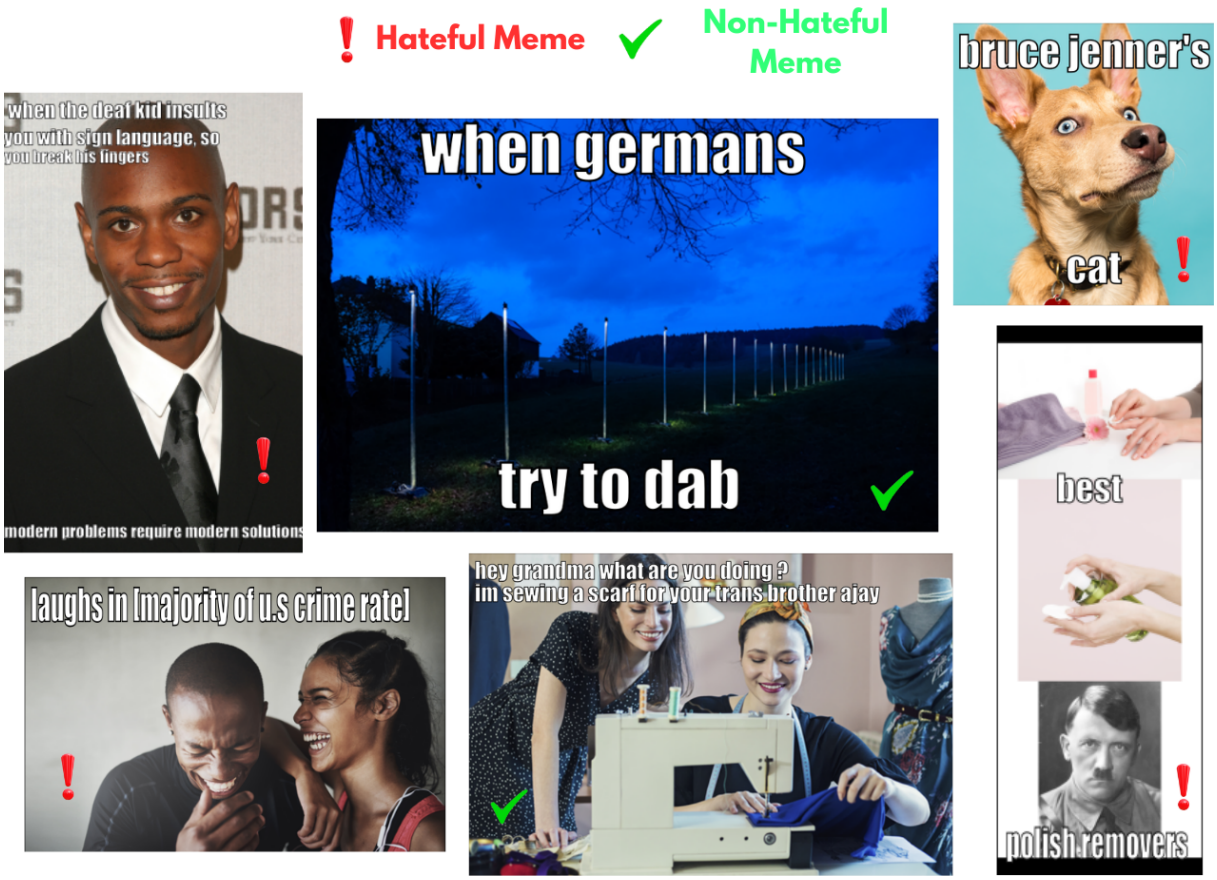
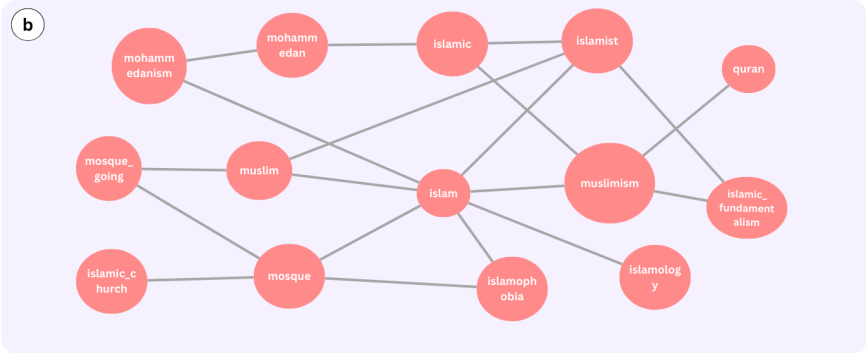


Figure 4: Examples from the Hateful Memes Dataset

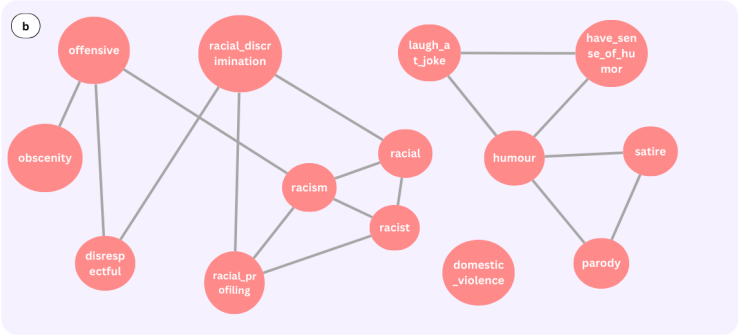
Figure 5: Examples from the HarMeme Dataset. The labels are given in the format [Intensity , Target] (Target label is not defined for not harmful memes)



c

LLaVA Caption : The meme is a two-panel image with a humorous and satirical tone, playing on the contrast between the desire to enter Islam and the reality of leaving it. In the top panel, there are two men embracing each other. The man on the left is wearing a white turban, which is often associated with Sikhism, and a red robe, which could be interpreted as a traditional South Asian outfit. The man on the right is wearing a white robe with a patterned design, which might suggest a religious or cultural attire. Both men are smiling and appear to be in a joyful or celebratory mood.The text above the image reads, "when you want to enter Islam." This suggests that the meme is targeting individuals who are interested in converting to Islam.In the bottom panel, the scene is dramatically different. A large crowd of people is shown, with a man lying on a stretcher in the center. The man is wearing a black and white striped shirt, which is reminiscent of the clothing worn by prisoners or those in a state of distress. The crowd around him is dense, with people of various ethnicities and genders. The expressions on their faces are not visible, but their presence suggests a sense of chaos or urgency.The text below the image reads, "when you want to leave Islam." This implies that the meme is also targeting individuals who have left Islam, suggesting that the process of leaving the religion is difficult or dangerous.The meme is using the contrast between the two panels to make a statement about the perceived challenges of converting to and leaving **Islam**. It is important to note that the meme is a form of satire and is not intended to be taken literally. It is also worth noting that the meme does not provide any specific details about the ethnicity, gender, or other characteristics of the individuals in the image, as these are not the focus of the meme.

Figure 6: ⓐ Example meme image from the Hateful Meme Dataset. ⓑ Portion of extracted subgraph after Relevancy Scoring and Pruning. ⓒ Generated LLaVA caption for the meme



c

LLaVA Caption : The meme features a close-up of a person with a bruised eye, which appears to be the result of a physical altercation. The individual has blonde hair and is looking directly at the camera with a serious expression. The text overlay on the image reads, "over heard my daughter telling her friend she was curious what it would be like to date a black guy so... I showed her!" The meme is a form of satire, using humor to critique or comment on societal issues. The primary subject of the meme is the stereotype and prejudice that can exist towards interracial relationships, particularly those involving Black individuals. The meme suggests that the person's daughter was curious about dating a Black person, and the parent's response was to physically harm her, presumably to discourage such a relationship. The meme is targeting the perpetuation of racial prejudice and the violence that can stem from it. It is important to note that the meme is not promoting violence but is using it as a metaphor to highlight the harm that can be caused by racism and discrimination. The meme is not specific to any particular group or individual, but rather addresses a broader societal issue.

Figure 7: ⓐ Example meme image from the Hateful Meme Dataset. ⓑ Portion of extracted subgraph after Relevancy Scoring and Pruning. ⓒ Generated LLaVA caption for the meme

Prompts for Dataset	Prompt for Teacher LVLM
Prompts for Captioning HateMeme	`You are given a meme. Write a detailed caption which captures the context, emotions and the theme of the meme. Identify and elaborate on the primary subject or target of the meme, which could include specific groups or individuals such as Muslims, Jews, transgender individuals, gays, Black people, Asians, etc.`
Prompts for Captioning HarMEME	`You are given a meme. Write a detailed caption that captures the context, emotions, and theme of the meme. Describe all visible elements in the image, including the ethnicity, gender, and other characteristics of individuals. Do not omit or censor any details.`

Table 8: Prompts used to generate captions and analyze meme content using the LLaVA model.