

Just KIDDIN' : Knowledge Infusion and Distillation for Detection of INdecent Memes

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

Rahul Garg* IIIT Hyderabad Hyderabad, India rahul.garg@research.iiit.ac.in	Trilok Padhi* Georgia State University Atlanta, GA, USA tpadhi1@student.gsu.edu	Hemang Jain IIIT Hyderabad Hyderabad, India hemang.jain@students.iiit.ac.in
--	---	---

Ugur Kursuncu
Georgia State University
Atlanta, GA, USA
ugur@gsu.edu

Ponnurangam Kumaraguru
IIIT Hyderabad
Hyderabad, India
pk.guru@iiit.ac.in

Abstract

Detecting toxicity in online multimodal environments, such as memes, remains a challenging task due to the complex contextual connections across modalities (e.g., text and visual), which demand both common-sense reasoning and contextual awareness. To bridge this gap, we propose a hybrid neurosymbolic framework that unifies (1) distillation of implicit contextual knowledge (e.g., sarcasm, cultural references) from Large Vision-Language Models (LVLMs) and (2) infusion of explicit relational semantics through sub-graphs from Knowledge Graphs (KGs). Experimental results on two benchmark datasets show the superior performance of our approach, *Knowledge-Infused Distilled Vision-Language Model (KID-VLM)*, over the state-of-the-art baselines across AUC and F1, with improvements of 0.5%, and 10.6%, respectively, in HatefulMemes Benchmark across variants. Further, KID-VLM demonstrates better generalizability and achieves the best performance across all baselines in the HarMeme Dataset with a 6.3% and 3.2% in F1 and AUC. Given the contextual complexity of the toxicity detection, KID-VLM showcases the significance of learning compact models (~500M parameters) from both explicit (i.e., KG) and implicit (i.e., LVLMs) contextual cues incorporated through a hybrid neurosymbolic approach. Our codes and pretrained models are publicly available¹.

1 Introduction

Online platforms have been increasingly prominent in disseminating harmful content, challenging the creation of safe digital spaces (Alatawi et al., 2021; Kursuncu et al., 2019b). According to the PEW

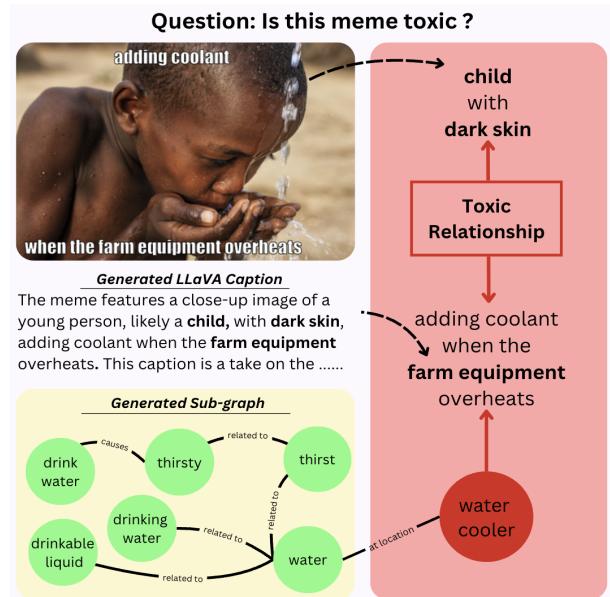


Figure 1: Given a meme, we aim to derive the answer by joint reasoning over the knowledge from LVLM, the KG (green box), and reason over toxicity (red box).

Research Center, 41% of Americans report experiencing online harassment (Vogels, 2021), and recent research highlights a surge in religious hate speech during the COVID-19 pandemic (Chandra et al., 2021). Detecting toxic content, especially nuanced and context-dependent multimodal forms, such as memes, remains a complex task (Kiela et al., 2020b; Sheth et al., 2022). Pavlopoulos et al. (2020) noted that context availability alters the perceived toxicity in 5.2% of posts. Memes that appear harmless can convey harmful messages through sarcasm, irony, or cultural references, thus potentially reinforcing negative stereotypes or societal norms (Waseem and Hovy, 2016; Ben Moussa et al., 2020; Schmid, 2025). Hence, accurate toxicity assessment requires a thorough understanding

*Equal contribution.

¹<https://github.com/SWAN-AI/Knowledge-Infused-Distilled-VLMs>

of context, as implicit evaluations can misinterpret tone or intent (Kursuncu et al., 2021; Sheth et al., 2022).

Although online multimodal toxicity detection has witnessed improvements, current methods, such as PromptHate (Cao et al., 2022a) and Hate-CLIPper (Kumar and Nandakumar, 2022a), rely solely on training data and pre-trained models, which limits their ability to capture complex contextual cues and nuances of toxicity. Moreover, while larger models, such as Flamingo (Alayrac et al., 2022) and LENS (Berrios et al., 2023), achieve comparable performance, they require substantial computational resources to deploy in limited-resource settings. These models often struggle to effectively address nuanced or context-dependent toxicity due to their reliance on pattern recognition without incorporating broader contextual knowledge (Pavlopoulos et al., 2022).

Our approach, *Knowledge-Infused Distilled Vision-Language Model* (*KID-VLM*), addresses these limitations by leveraging ConceptNet (Speer et al., 2017), a common sense knowledge graph (KG) that enhances the comprehension of contextual nuances. In this study, the following research questions guide our investigation:

RQ1: How can we incorporate Knowledge Distillation (KD) from LVLMs and common sense knowledge from KGs into compact models, enhancing their effectiveness in detecting toxicity within memes?

RQ2: How does the generalizability of KID-VLM compare to baseline methods in detecting toxicity across datasets? *HatefulMemes* (Kiela et al., 2020b) and *HarMeme* (Pramanick et al., 2021b)?

We present a novel approach that enhances *compact* VLMs by combining KD from LVLMs with knowledge infusion from KGs, enriching them with both implicit and explicit contextual knowledge (see Figure 1). Utilizing CLIP (Radford et al., 2021) as the backbone, we extract visual and textual features from memes, supplemented with captions generated by the LLaVA 1.6-NeXT model (Liu et al., 2024b) for additional contextual features. These features are then fused using Align Fusion (Kumar and Nandakumar, 2022b) to create a multimodal representation, refined with knowledge distilled from the teacher model’s caption representations through consistency loss. This process enables the student model to learn implicit contextual cues. Using graph-based reasoning, we fur-

ther incorporate external knowledge from ConceptNet, constructing a joint graph that combines the meme context with KG entities. This knowledge-enhanced multimodal representation is then optimized using cross-entropy loss for final toxicity predictions (see Figure 2).

Our findings demonstrate the effectiveness of the *KID-VLM* framework, which outperforms the baselines. Error analysis and ablation studies further highlight the significance of explicit and implicit context incorporated through KD and infusion, providing improvements of 10.6% and 0.5% in F1 and AUC, respectively. Further, our approach demonstrates better generalizability compared to baselines, as seen in our performance on the HarMeme Dataset, where KID-VLM outperforms all baselines with an AUC of 92.98.

2 Related Work

2.1 Online Toxicity Detection

Early approaches to online toxicity detection primarily focused on textual content, employing traditional machine learning techniques (Leo et al., 2023; Saha et al., 2023). These techniques have evolved to utilize deep learning in addressing its nuanced, context-dependent nature (Jonathan and Setiawan, 2023; Karim et al., 2022). Kursuncu et al. (2019a) incorporated multiple dimensions of online content, utilizing domain-specific corpora in detecting malicious actors. Multimodal Models emerged (Kumar and Nandakumar, 2022a),(Cao et al., 2022a), leveraging pre-trained VLMs and prompt tuning to improve performance. However, these methods lack mechanisms to integrate explicit, structured reasoning, such as socio-cultural norms, necessary for identifying subtle forms of hate in memes. Large multimodal systems, including Flamingo-80B (Alayrac et al., 2022) and LENS (Berrios et al., 2023), demand high computational costs, challenging deployment (Erol et al., 2025). Recent work, such as Pro-Cap (Cao et al., 2023), has attempted to address this by freezing VLMs and aligning textual-visual features for efficient inference. However, it still solely relies on training data and retains dependence on monolithic models without external knowledge. Efforts to improve interpretability included (?), which distilled reasoning from LLMs into smaller architectures and a multimodal debate framework (Lin et al., 2024) that enhances transparency in meme analysis.

2.2 Knowledge Enhanced Learning

Knowledge Enhanced Learning can be broadly categorized into two complementary approaches: (i) Knowledge Infusion (KI), which integrates structured external knowledge, such as KGs and ontologies, into models, and (ii) Knowledge Distillation (KD), where a large teacher model guides the training of a compact student model, helping to achieve similar performance with lower computational cost.

KI from KGs. KI enhances the model’s representations by incorporating external structured knowledge from KGs and ontologies to capture explicit context across modalities (Zhang et al., 2022; Kurşuncu et al., 2020; Khandelwal et al., 2024). Xu et al. (2024) have improved synthetic clinical data generation by combining KGs with LLMs, enhancing performance, and addressing privacy and fairness concerns. Lymperaiou and Stamou (2024) demonstrated how KI can improve common sense and temporal reasoning in multimodal contexts. Further, Agrawal et al. (2023) investigated how KGs can reduce hallucinations in LLMs, while Padhi et al. (2024) enhanced cross-modal contextual congruence for crowdfunding campaigns.

KD from Large Models. KD is a process in which knowledge is transferred from a larger teacher model to a smaller, more efficient student model. Wang et al. (2020) demonstrated that KD can effectively address incomplete modalities by distilling information from a fully equipped multimodal system. Building on this idea, Padhi et al. (2025) demonstrated that external knowledge feedback enhances grounding, thereby improving calibration of multimodal LLMs. This highlights how incorporating cross-modal signals enhances model confidence and robustness. Similarly, Hong and Zhen (2023) applied KD to Graph Neural Networks (GNNs) to enhance the capturing of community structures. Setiawan (2024) introduced an n -best reranking method to refine sequence-level KD for neural machine translation. In contrast, Gholami et al. (2024) proposed an out-of-distribution-guided framework that enhances distilled models through targeted data generation. Further, Wei et al. (2024) compared sentence-level and token-level distillation under optimal conditions for each approach.

In contrast to prior work, our approach unifies KD with KI via a neurosymbolic framework that infuses explicit relational semantics from commonsense KGs and distills implicit knowledge from

LVLMs. This approach aims to equip compact models with the ability to learn from both knowledge bases and large pre-trained models, thereby boosting their multimodal reasoning capabilities.

3 Methodology

Our approach, KID-VLM, illustrated in Figure 2, utilizes a frozen Vision-Language (VL) Encoder as the student model S , ConceptNet as the KG K , and an LVLM as teacher model T , serving as non-parametric knowledge sources to query explicit and implicit knowledge, respectively. For each data point, a meme with an image I_i and overlaid text T_i , a joint multimodal feature representation s_i is generated using the pre-trained student model, $s_i = S(I_i, T_i)$. To get the joint representation, we use the Align Fusion for it’s low computational resource conditions (Kumar and Nandakumar, 2022b). Concurrently, the teacher model T produces a caption C_i capturing the meme’s implicit context, which is used for two purposes: (i) aid in teaching the student model through KD loss \mathcal{L}_{KD} , and (ii) query ConceptNet to extract relevant sub-KGs G_{sub} for each data point. The distilled representation $h_{distilled}$ from the student model is then fused with the pooled representation h_{graph} from sub-KGs obtained through a GNN G , $h_{graph} = G(G_{sub})$. This fused representation is optimized through Cross Entropy loss L_{CE} , enabling the model to synthesize outputs from both the teacher’s latent reasoning and the explicit relational knowledge from the KG.

3.1 Knowledge Distillation

We utilize HateCLIPper (Kumar and Nandakumar, 2022b) as our frozen VL Encoder for stable feature extraction from each meme, while the teacher model, LLaVA-NeXT (Liu et al., 2024b), generates captions that capture the implicit context of the memes (See Appendix Table 12) KD is employed only during the training stage to optimize the student model’s multimodal representation, s_i , by minimizing the consistency loss \mathcal{L}_{KD} , defined as the Euclidean distance between the student’s s_i and the representations of the captions obtained from the teacher model. \mathbf{w}_i^{LLaVa} :

$$\mathcal{L}_{KD} = \|\mathbf{s}_i - \mathbf{w}_i^{LLaVa}\|_2^2 \quad (1)$$

This KD process, which does not extend to the testing stage, aligns the internal features of the

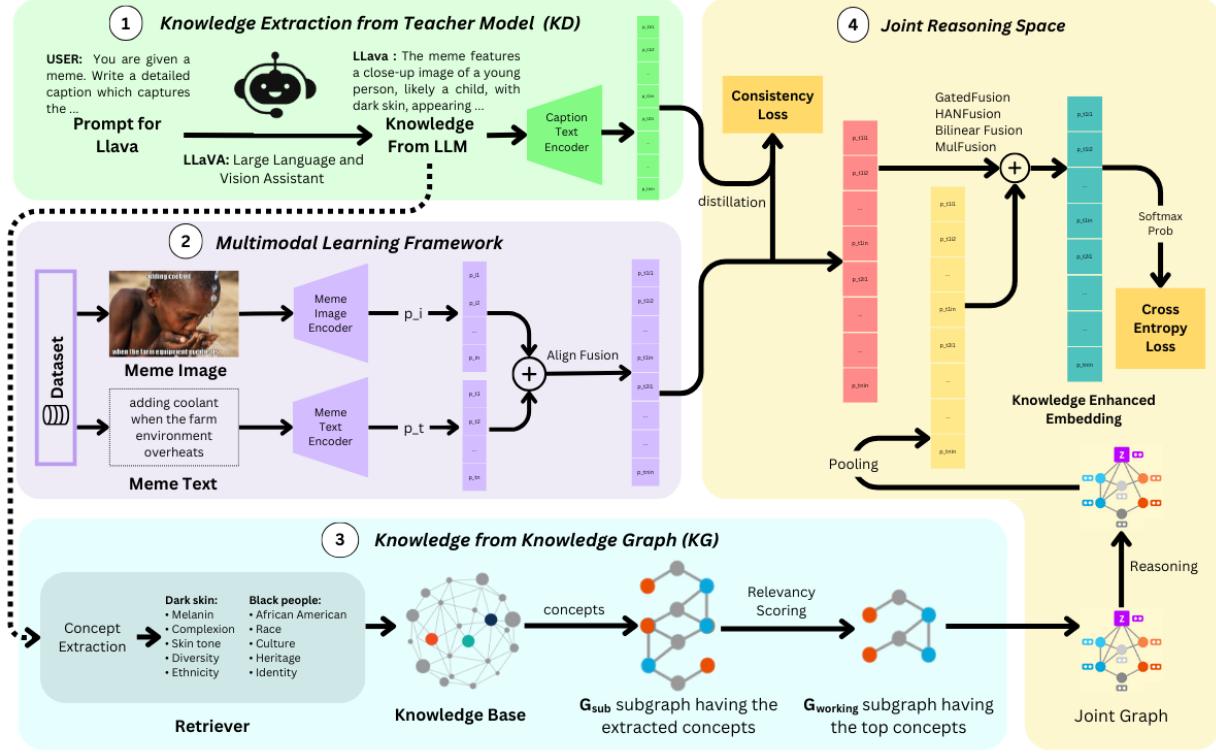


Figure 2: **KID-VLM** framework: The framework unifies KD from an LVLM with KI from external KGs such as ConceptNet. The input image and text are processed through the CLIP encoders to generate embeddings, which are fused using different fusion mechanisms. ① Knowledge Extraction from Teacher Model. ② 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.

student model with those of the teacher model, enabling the student to emulate the teacher’s latent reasoning and recognize subtle implicit contextual cues in the memes. Note that LLaVA is not used at inference time. It is solely used in preprocessing to create captions that guide knowledge retrieval, ensuring our compact student model maintains a lightweight inference footprint and enabling deployment in low-resource settings.

3.2 Graph-Based Reasoning

To enrich the distilled multimodal representation with explicit relational knowledge, we employ a joint reasoning approach, which constructs joint *working graphs* G_{working} for each data point, providing a reasoning space for the meme’s overlaid text T_i , sub-KG, and the generated caption C_i . Then, this process generates a knowledge representation for working graphs through GNNs (Yasunaga et al., 2021).

3.2.1 Working Graph Construction

To construct the working graph G_{working} , we first extract a sub-KG G_{sub} from ConceptNet (See Fig-

ure 2) incorporating concepts from the meme’s text and the teacher-generated caption. We explicitly connect this sub-KG in a joint graph with a new context node z , which represents the meme context and connects to each entity from the meme’s text and teacher-generated text. The heuristic retrieval of G_{sub} from a KG can introduce irrelevant nodes, potentially adding noise and complicating reasoning, especially with large sub-KGs (Yasunaga et al., 2021).

To assess the semantic alignment between the context node z and each candidate node $v \in G_{\text{sub}}$, we apply two relevance scoring methods. The first approach uses the RoBERTa language model (Liu et al., 2019), where the perplexity score of the concatenated textual inputs $[z; v]$ is used to evaluate relevance. The second utilizes MiniLM (Wang et al., 2021) in conjunction with Sentence Transformers to compute the cosine similarity between the contextual and entity embeddings. For all main experiments and benchmark evaluations, we use MiniLM-based scoring due to its computational efficiency and strong performance. However, to evaluate robustness, we additionally report results

using RoBERTa-based scoring in a separate ablation study (see Table 3).

The nodes within this sub-KG are ranked by relevance scoring to minimize noise and focus on the most relevant KG entities. This scoring involves evaluating the relationship of each node $v \in V_{\text{sub}}$, where V_{sub} represents the nodes of G_{sub} , with the overall meme context, encapsulated by the context node z . The relevance score ρ_v is calculated through the Roberta model (Liu et al., 2019) by passing the concatenated text representations of z and v . The relevance score is determined using the perplexity score:

$$\rho_v = f_{\text{head}}(f_{\text{enc}}([\text{text}(z); \text{text}(v)])) \quad (2)$$

where f_{enc} represents the encoder (Roberta) and f_{head} is the scoring head. A lower perplexity score indicates higher semantic relevance, helping prioritize nodes that align closely with the meme’s context. For additional validation, we compute the cosine similarity between the concatenated text representation of z and KG entities v using MiniLM (Wang et al., 2021) (Sentence-Transformer) as the relevance score $\rho_v = \text{cosine}(z, v)$ to help identify the contextually significant nodes. We separately assess the relevance using scores from both Roberta and MiniLM in our experiments, selecting the top $k = 750$ most relevant entities to include in the working graph G_W (referred to as G_{working} in Figure 2). After constructing G_W , context node z is added to bridge knowledge from the teacher model with that from KG. This node is then linked to the entities V_{sub} in the working graph with a new relationship r' , finalizing G_W for subsequent knowledge representation learning.

3.2.2 Knowledge Representation Learning

To process and learn the representation of the joint working graph G_W , we employ the *Relational Graph Convolutional Network (R-GCN)* (Schlichtkrull et al., 2017), which enhances node representations by applying relation-specific transformations to aggregate information from neighboring nodes through structured interactions:

$$h_v^{(l+1)} = \sigma \left(\sum_{r \in R} \sum_{u \in \mathcal{N}_r(v)} W_r^{(l)} h_u^{(l)} \right) \quad (3)$$

where $h_v^{(l+1)}$ is the updated representation of node v at layer $l + 1$, σ is the activation function, $W_r^{(l)}$ is the relation-specific transformation matrix, $\mathcal{N}_r(v)$ represents the set of neighbors connected to v by

relation r , and $h_u^{(l)}$ is the representation of a neighboring node u at layer l . In addition, we learn a pooled representation of the graph to generate a single vector representing the entire working graph using Mean Pooling (Appendix A).

Fusion Mechanism: The pooled graph representation h_{graph} is then fused with the distilled multimodal representation $h_{\text{distilled}}$ through *Gated Fusion* (Xiong et al., 2024) mechanism:

$$\mathbf{F}_{\text{multimodal}} = \mathbf{G} \odot h_{\text{graph}} + (1 - \mathbf{G}) \odot h_{\text{distilled}} \quad (4)$$

where $\mathbf{F}_{\text{multimodal}}$ is fused multimodal representation and \mathbf{G} is the gating mechanism given by:

$$\mathbf{G} = \sigma(\mathbf{W}_g (h_{\text{graph}} \| h_{\text{distilled}})) \quad (5)$$

\mathbf{W}_g is a learnable weight matrix, and σ is the sigmoid activation function.

Loss Function: The model is trained using a joint loss function:

$$\mathcal{L}_{\text{total}} = \lambda_1 \mathcal{L}_{\text{BCE}} + \lambda_2 \mathcal{L}_{\text{KD}} \quad (6)$$

where \mathcal{L}_{BCE} is the Binary Cross-Entropy Loss for meme classification, and \mathcal{L}_{KD} is the Consistency Loss for distillation, ensuring alignment between the multimodal fusion and captions from the teacher model. The losses are weighted using hyperparameters, and the overall loss is minimized using the AdamW optimizer.

3.3 Experimentation and Evaluation

We conducted experiments on two benchmark datasets: HatefulMemes (Kiela et al., 2020b) and HarMeme (Pramanick et al., 2021a). Detailed statistics are provided in Appendix B. For evaluation consistency with prior research (Kumar and Nandakumar, 2022a; Cao et al., 2022a), we used AUC, F1 score, Precision, Recall, and Accuracy. The HatefulMemes dataset provided two distinct ‘Seen’ and ‘Unseen’ data splits to assess the generalization ability and robustness of each model. The datasets include text annotations for memes, utilized as the text modality, avoiding any influence from LVLM’s OCR limitations.

KI: We employed ConceptNet for explicit contextual information. We leveraged multi-hop traversal to expand the model’s contextual understanding, specifically experimenting with *Hop 1* and *Hop*

Framework	Accuracy		F1		Precision		Recall		AUC	
	Seen	Unseen	Seen	Unseen	Seen	Unseen	Seen	Unseen	Seen	Unseen
MMBT	68.80	70.85	67.61	64.73	71.25	70.55	68.44	64.47	77.84	77.12
CLIP*	59.50	70.85	53.88	65.67	76.40	69.75	25.10	65.21	72.14	76.25
CLIP	60.30	62.80	57.10	55.50	69.62	58.45	33.67	56.19	68.30	57.78
LLaVa ⁺	54.30	63.55	29.58	19.60	60.38	53.85	19.59	28.74	-	-
PromptHate	76.10	75.75	75.76	72.94	77.30	74.66	76.10	72.20	84.21	80.43
ExplainHM + LLM Debate	69.40	68.05	65.23	55.72	73.59	58.01	58.57	53.50	74.61	71.82
Pro-Cap + Targeted Captions	75.50	74.35	75.30	73.46	76.02	73.21	75.35	74.41	83.65	81.34
HateClipper	76.07	76.65	73.00	64.21	78.00	75.50	56.00	55.87	85.14	83.70
RGCL	78.30	77.65	76.23	67.20	82.27	74.71	71.02	61.07	86.72	84.46
KID-VLM + Hop 1 & MiniLM	78.70	77.00	78.63	73.87	80.57	76.69	74.49	72.91	86.11	84.86
KID-VLM + Hop 2 & MiniLM	76.30	76.85	75.95	74.29	82.52	75.84	65.51	73.56	85.48	83.99

Table 1: Performance Comparison of various models on the *HatefulMemes Dataset* (seen & unseen splits). Top values for each metric are expressed in (**bold**). * represents Frozen CLIP encoders, + represents zero-shot evaluation.

2 expansions. We also varied the top_k most relevant nodes, where k ranged from 250 to 750. This range was chosen because higher values exceeded resource limitations, while fewer than 250 nodes compromised the depth of contextual information. **Fusion:** For integrating multimodal and graph representations, we explored several fusion mechanisms, including *Gated Fusion*, *Multiplicative Fusion*, *Bilinear Fusion*, and *Hierarchical Attention Fusion* (Appendix C). To process the G_w , we utilized *RGCN* for all our experiments based on our ablation study results.

3.4 Baselines

To evaluate the performance of our KID-VLM approach, we benchmarked against several *compact* VLMs from prior research, which include HateClipper (Kumar and Nandakumar, 2022b), RGCL (Mei et al., 2023), Pro-Cap (Cao et al., 2023), Explain-HM (Lin et al., 2024), Mr. Harm (Lin et al., 2023), and PromptHate (Cao et al., 2022b). We also incorporated the Multimodal Bitransformers (MMBT) (Kiela et al., 2020a) and various implementations of CLIP-based models (Radford et al., 2021) to facilitate a broad comparative analysis. In addition, we fine-tuned only the projection layers of CLIP, designated as CLIP*. The LLaVA-NeXT (Liu et al., 2024b) was used in a zero-shot setting to compare with a Large VLM. Our primary baseline reference for all reported values is RGCL.

3.5 Implementation Details

We utilized Optuna to optimize the model’s hyperparameters. Our parameter sweep included GNN

hidden dimensions, number of layers, learning rate, fusion methods, and dropout rates. GNN hidden and output dimensions ranged from 2 to 512 and 2 to 1024, respectively. Mapping layer dimensions were set between 2 and 2048. Learning rates ranged from $1e-10$ to $1e-2$, and dropout probabilities from 0.0 to 0.9. We also tuned the number of mapping and pre-output layers (1 to 5), weight decay ($1e-8$ to $1e-1$), and the loss alpha (0 to 1). The model was trained for 30 epochs via the AdamW optimizer, with learning rate warm-up and linear decay. We used a batch size of 4 on RTX 5000 and L40 GPUs with 32 and 40 GB of VRAM, respectively. Model performance was monitored on the validation set, and the best model was selected based on validation AUC. The total trainable parameter count in KID-VLM remains well under 1B, including the CLIP encoders, the graph reasoning modules (R-GCN), and fusion layers, aligning with the compact model threshold from prior work (Izsak et al., 2019; Houlsby et al., 2019), which positions KID-VLM as performant and efficient.

4 Results & Discussion

Our analysis shows that *KID-VLM*, which integrates external knowledge from ConceptNet and distills information from the LLaVA model, outperforms the competitive baseline models from prior research (Mei et al., 2023; Cao et al., 2023; Lin et al., 2024; Cao et al., 2022b) across both the *HatefulMemes* and *HarMeme* datasets (see Tables 1 and 2). On the *HatefulMemes* dataset across variants, our approach outperforms by 3.2% in F1 score for

Framework	Accuracy	F1	Precision	Recall	AUC
CLIP*	70.62	58.01	73.79	59.55	90.80
CLIP	66.67	55.04	62.21	56.51	80.82
LLaVa ⁺	80.23	67.59	79.35	58.87	-
PromptHate	78.53	78.28	80.02	82.73	90.72
Pro-Cap+Target-based Captions	83.90	83.18	82.62	85.38	90.97
ExplainHM+LLM Debate	84.75	77.50	80.17	75.00	92.05
Mr.Harm+Label Rationale	74.29	62.76	85.83	63.31	-
HateClipper	83.33	79.29	70.19	91.13	91.90
RGCL	84.75	79.39	75.36	83.87	90.10
KID-VLM+Hop 1 & MiniLM	84.46	83.85	83.38	86.37	92.77
KID-VLM+Hop 2 & MiniLM	85.03	84.40	83.86	86.81	92.98

Table 2: Performance Comparison of various models on HarMeme Dataset across multiple metrics. The top values for each metric are expressed in (**bold**). * represents Frozen CLIP encoders, + represents zeroshot evaluation.

the seen split and 10.6% for the unseen split. It also improves AUC by 0.5% on the unseen split. For the *HarMeme* dataset, it shows improvements of 6.3% and 3.2% in F1 and AUC, respectively. The KID-VLM with MiniLM and Hop 2 traversal demonstrate overall higher performance on both datasets. KID-VLM with Hop 1 models also perform competitively, often matching or outperforming other models. We observe that baseline models often exhibit a trade-off between precision and recall, whereas KID-VLM models show a notable balance, reflected in higher F1 and AUC scores. This robust performance across metrics suggests the ability to capture the underlying patterns of toxicity. Further, our models’ performance on the unseen split of the HatefulMemes dataset highlights better generalization capabilities, suggesting that the broader contextual understanding afforded by Hop 2 allows the model to generalize beyond the training data.

4.1 Performance on HatefulMemes Dataset

Our framework, utilizing both Hop 1 and Hop 2 traversals, demonstrates consistent robust performance on both seen and unseen splits of the *HatefulMemes* dataset, outperforming baseline models. Specifically, the KID-VLM with Hop 1 and Gated Fusion achieved an F1 score of 78.63, an AUC of 86.11, and an accuracy of 78.70. Notably, the Hop 2 variant achieved the highest recall at 76.14 while maintaining robust accuracy and AUC, highlighting its capability to detect contextual nuances of toxicity. The KID-VLM models showcase a well-balanced trade-off between precision and recall. The integration of *MiniLM* for relevancy scoring in the creation of a working graph improved per-

formance, which can be attributed to MiniLM’s ability to refine entity selection, leading to a more semantically coherent working graph, enhancing graph-based reasoning by reducing sparsity, and improving relational inference. While MMBT, CLIP, and LLaVA show varied results, CLIP-based models underperform on the unseen split, particularly in precision and recall. HateClipper and RGCL provide competitive baselines, with RGCL displaying an AUC of 86.72 on the seen split.

4.2 Performance on HarMeme Dataset

On the *HarMeme* dataset, KID-VLM with Hop 2 achieved the highest overall performance with an F1 score of 84.40 and an AUC of 92.98, which represents the state-of-the-art for this dataset. The Hop 1 model also outperforms the baselines, demonstrating the efficacy of the KID-VLM framework integrating external knowledge for meme classification. The improvement between Hop 2 and Hop 1 indicates the capturing of broader contextual cues enhancing performance. Mr.Harm and HateClipper outperform in precision and recall, respectively, while both fall short in recall and precision, showcasing a trade-off. Pro-Cap and ExplainHM exhibit more balanced results, albeit not matching the overall performance of KID-VLM. RGCL notably underperforms on this dataset, in contrast to its performance on the seen split of the HatefulMeme dataset (see Table 1).

4.3 Impact of Node Count

The ablation study, summarized in Table 4, illustrates the impact of increasing node count from KG on the model performance in AUC. With a node count of 250, the model achieves an AUC of 84.91,

KID-VLM Variant (RoBERTa)	Accuracy		F1		Precision		Recall		AUC	
	Seen	Unseen	Seen	Unseen	Seen	Unseen	Seen	Unseen	Seen	Unseen
KID-VLM + Hop 1 & RoBERTa	78.20	76.85	78.12	74.16	80.36	75.96	73.47	73.37	86.02	83.50
KID-VLM + Hop 2 & RoBERTa	76.30	77.00	76.07	73.58	76.94	77.16	76.14	72.53	86.11	84.72

Table 3: Performance of **KID-VLM variants using RoBERTa-based relevance scoring** on the *HatefulMemes Dataset* (seen & unseen splits). These results complement the MiniLM-based variants in Table 1.

which improves to 85.76 with 500 nodes. When the node count is raised to 750, the model yields the highest AUC of 86.11. This trend indicates that incorporating more knowledge in the model from ConceptNet enhances its ability to discern and utilize contextual relationships within memes, thereby improving its reasoning capabilities and overall performance. On the other hand, it is crucial to ensure the relevance of the nodes to maintain performance quality and avoid introducing noise.

#	Nodes	AUC	#	Fusion	AUC
1	250	84.91	1	Gated	86.11
2	500	85.76	2	Multiplicative	80.64
3	750	86.11	3	Bilinear	85.26
-	-	-	4	Hierarchical Att.	85.39

Table 4: Ablation studies on KID-VLM for the Hateful-Memes Dataset analyze node count and fusion methods. The left section shows that increasing nodes (250, 500, 750) with Hop-1, RGCN, and Gated Fusion improves performance. The right section evaluates fusion mechanisms with Hop-2, 750 nodes, and CLIP, identifying Gated Fusion as the most effective based on AUC.

4.4 Fusion & GNN Architecture

Our ablation study (Table 4) demonstrates the effectiveness of *Gated Fusion*, balancing multimodal and graph-based features to prevent any modality from biasing the fusion process. Gated Fusion utilizes a dynamic gating mechanism that adaptively weighs the information from multimodal and graph-based representations using a sigmoid-activated gate. This allows for minimizing noise from external knowledge and preserving contextually relevant information, essential for toxicity detection. Gated fusion outperforms others, such as multiplicative fusion, which may diminish key signals in sparse feature space, and bilinear fusion, which may introduce complexity and increase the risk of overfitting. Hierarchical Attention may risk overlooking critical subtle cues necessary for identifying nuanced toxicity. Further, our experiments with GCN and GAT suggest that GCN, with an AUC of 86.11, provides better performance through uniform fea-

ture aggregation. In contrast, GAT, with an AUC of 85.75, may be more prone to noise due to its attention mechanism (Refer Appendix A).

Framework	Acc.	F1	Pre.	Rec.	AUC
No Knowledge	76.07	73.00	78.00	56.00	85.14
KI-VLM	78.10	77.63	78.09	77.55	85.35
KD-VLM	73.70	73.12	75.33	73.44	85.53
KID-VLM	78.70	78.63	80.57	74.49	86.11

Table 5: Ablation study on the impact of KI and KD on HatefulMemes Dataset using Hop-1 (750 nodes) and MiniLM for Relevancy Scoring.

4.5 Impact of KI and KD

The ablation study shown in Table 5 demonstrates the complementary benefits of KI and KD in enhancing toxicity detection. KI-VLM leverages external commonsense knowledge from ConceptNet, improving Recall by 38.48%, F1 score by 6.34, and AUC by 0.25% over the baseline VLM with no knowledge, highlighting the value of explicit contextual reasoning. Similarly, KD-VLM enhances Recall by 31.14% and AUC by 0.46%, suggesting that distilled knowledge from LVLMs improves implicit contextual understanding. The KID-VLM framework, which combines both KI and KD, achieves the best overall performance with improvements in AUC of 1.14%, F1 score of 7.71%, and accuracy of 3.46%, showcasing the effectiveness of fusing explicit and implicit knowledge for robust multimodal toxicity detection. In addition to MiniLM-based relevance scoring (Table 1), we included results using RoBERTa-based scoring in Table 3, showing comparable performance across splits, with the Hop 2 variant demonstrating higher recall (76.14) on seen data and AUC of 84.72 on unseen data.

4.6 Discussion

Better Separation: Toxic vs. Non-Toxic. Enhancing representations by KI leads to a more distinct separation between toxic and non-toxic content as shown in Figure 3. Without knowledge-enhanced representations, the separation between

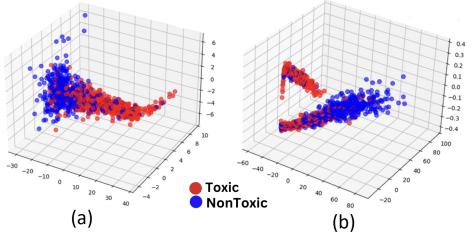


Figure 3: Baseline (without KI/KD) vs. KID-VLM (with KI/KD) t-SNE plots illustrating the reduced 3D representation of the dataset after dimensionality reduction. The colors represent the ground truth labels of the data points. KID-VLM’s plot shows a much clearer separation between the labeled data points.

these two classes tends to be ambiguous, and the data points are more scattered in the latent space. In contrast, with knowledge, there is a clearer separation through distinct regions with denser clusters. This improved separation reduces the likelihood of misclassifications, which is often caused by ambiguity and unclear separation, especially in borderline cases where the context is crucial for detecting toxicity in a meme. This clearer separation is likely to stem from both explicit and implicit contextual cues learned during the training.

Efficient Training and Deployment. KID-VLM minimizes training overhead by using LVLMs only during inference, avoiding large-scale multimodal model training in prior works (Lin et al., 2024; Liu et al., 2024a). By distilling crucial knowledge into a compact model with ~500M parameters, rather than fine-tuning a large model, our approach substantially lowers computational costs. This facilitates efficient deployment in real-world settings where compute resources are limited, making KID-VLM scalable for toxicity detection applications.

Multi-Hop and Enhanced Generalization. Leveraging multi-hop traversal helps our model to capture extended contextual cues, improving the overall performance on both the HarMeme and HatefulMemes datasets. It achieves particularly high scores in F1 score and AUC for unseen data, indicating a robust ability to handle new and diverse content. This enhanced generalization likely stems from the broader contextual understanding that Hop 2 provides, which is essential for navigating complex, nuanced content.

5 Conclusion

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. By integrating explicit relational knowledge with implicit contextual cues, KID-VLM better captures sarcasm, cultural references, and multimodal nuances that often obscure toxicity. The use of *ConceptNet subgraphs* and distilled multimodal representations allows the model to understand the complex semantics of multimodal content better, resulting in improved accuracy, F1, and AUC scores. By distilling knowledge into a compact model (~500M parameters), KID-VLM achieves these gains without the computational burden of large-scale models, making it scalable for real-world deployment.

6 Limitations

While *KID-VLM* demonstrates strong performance on the *HatefulMemes* and *HarMeme* datasets, it has several limitations. The model’s reliance on ConceptNet may limit its generalizability to other datasets beyond the two that have been extensively examined in this study. Additionally, incorporating graph-based methods may increase computational complexity, which can affect scalability for larger datasets. The quality of KD from larger models could degrade when using smaller student models, and there remains a potential risk of bias from pre-trained models and KGs. Additionally, KID-VLM may inherit hallucination issues from its use of LLaVA. We have presented and discussed a few failure cases in Appendix E. Future work could address these concerns by exploring more diverse datasets, improving scalability, and investigating bias mitigation strategies.

7 Ethical Considerations

Toxicity detection systems can mislabel content or users, particularly when processing complex inputs such as memes that involve irony or satire. The cultural and social context may not be fully understood by the model, so continuous refinement is essential to mitigate biases. The study utilizes publicly available, anonymized datasets; however, the 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

- Garima Agrawal, Tharindu Kumarage, Zeyad Alghamdi, and Huan Liu. 2023. Can knowledge graphs reduce hallucinations in llms?: A survey. *arXiv preprint arXiv:2311.07914*.
- Faisal Alatawi, Lu Cheng, Anique Tahir, Mansooreh Karami, Bohan Jiang, Tyler Black, and Huan Liu. 2021. A survey on echo chambers on social media: Description, detection and mitigation. *arXiv preprint arXiv:2112.05084*.
- 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.
- Mohamed Ben Moussa, Sanaa Benmessaoud, and Aziz Douai. 2020. Internet memes as “tactical” social action: A multimodal critical discourse analysis approach. *International Journal of Communication*, 14:5920–5940.
- William Berrios, Gautam Mittal, Tristan Thrush, Douwe Kiela, and Amanpreet Singh. 2023. Towards language models that can see: Computer vision through the lens of natural language. *arXiv preprint arXiv:2306.16410*.
- Rui Cao, Ming Shan Hee, Adriel Kuek, Wen-Haw Chong, Roy Ka-Wei Lee, and Jing Jiang. 2023. Procap: 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, Wei Chong, and Jing Jiang. 2022a. Prompting for multimodal hateful meme classification. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (EMNLP’22)*.
- 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, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.
- 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.
- Abdulkadir Erol, Trilok Padhi, Agnik Saha, Ugur Kursuncu, and Mehmet Emin Aktas. 2025. Playing devil’s advocate: Unmasking toxicity and vulnerabilities in large vision-language models. *arXiv preprint arXiv:2501.09039*.
- Mohsen Gholami, Mohammad Akbari, Tianxi Hu, Vaden Masrani, Z. Wang, and Yong Zhang. 2024. Gold: Generalized knowledge distillation via out-of-distribution-guided language data generation. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 4365–4380. Association for Computational Linguistics.
- Tang Hong and Liu Zhen. 2023. Research on Community Discovery Techniques Based on Knowledge Distillation. 2023 20th International Computer Conference on Wavelet Active Media Technology and Information Processing (ICCWAMTIP), pages 1–4.
- Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. 2019. Parameter-efficient transfer learning for nlp. In *International conference on machine learning*, pages 2790–2799. PMLR.
- Peter Izsak, Shira Guskin, and Moshe Wasserblat. 2019. Training compact models for low resource entity tagging using pre-trained language models. In *2019 Fifth Workshop on Energy Efficient Machine Learning and Cognitive Computing-NeurIPS Edition (EMC2-NIPS)*, pages 44–47. IEEE.
- 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 Shahjalal, and Bharathi Raja Chakravarthi. 2022. Multimodal Hate Speech Detection from Bengali Memes and Texts. ArXiv:2204.10196 [cs].
- Vedant Khandelwal, Manas Gaur, Ugur Kursuncu, Valérie L Shalin, and Amit P Sheth. 2024. A domain-agnostic neurosymbolic approach for big social data analysis: Evaluating mental health sentiment on social media during covid-19. In *2024 IEEE International Conference on Big Data (BigData)*, pages 959–968. IEEE.
- 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*.
- Ugur Kursuncu, Manas Gaur, Carlos Castillo, Amanuel Alambo, Krishnaprasad Thirunarayan, Valerie Shalin, Dilshod Achilov, I Budak Arpinar, and Amit Sheth. 2019a. Modeling islamist extremist communications on social media using contextual dimensions: religion, ideology, and hate. *Proceedings of the ACM on Human-Computer Interaction*, 3(CSCW):1–22.
- Ugur Kursuncu, Manas Gaur, Usha Lokala, Krishnaprasad Thirunarayan, Amit Sheth, and I Budak Arpinar. 2019b. Predictive analysis on twitter: Techniques and applications. *Emerging research challenges and opportunities in computational social network analysis and mining*, pages 67–104.
- Ugur Kursuncu, Manas Gaur, and Amit Sheth. 2020. Knowledge infused learning (k-il): Towards deep incorporation of knowledge in deep learning. *Proceedings of the AAAI 2020 Spring Symposium on Combining Machine Learning and Knowledge Engineering in Practice (AAAI-MAKE 2020)*.
- Ugur Kursuncu, Hemant Purohit, Nitin Agarwal, and Amit Sheth. 2021. When the bad is good and the good is bad: understanding cyber social health through online behavioral change. *IEEE Internet Computing*, 25(1):6–11.
- 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.
- Hongzhan Lin, Ziyang Luo, Wei Gao, Jing Ma, Bo Wang, and Ruichao Yang. 2024. Towards explainable harmful meme detection through multimodal debate between large language models. In *Proceedings of the ACM Web Conference 2024*, pages 2359–2370.
- Hongzhan Lin, Ziyang Luo, Jing Ma, and Long Chen. 2023. *Beneath the surface: Unveiling harmful memes with multimodal reasoning distilled from large language models*. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 9114–9128, Singapore. Association for Computational Linguistics.
- Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2024a. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26296–26306.
- Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. 2024b. *Llava-next: Improved reasoning, ocr, and world knowledge*.
- 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.
- Maria Lymperaiou and Giorgos Stamou. 2024. A survey on knowledge-enhanced multimodal learning. *Artificial Intelligence Review*, 57(10):284.
- Jingbiao Mei, Jinghong Chen, Weizhe Lin, Bill Byrne, and Marcus Tomalin. 2023. *Improving hateful meme detection through retrieval-guided contrastive learning*. *ACL 2024 Main*.
- Trilok Padhi, Ramneet Kaur, Adam D Cobb, Manoj Acharya, Anirban Roy, Colin Samplawski, Brian Matejek, Alexander M Berenbeim, Nathaniel D Bastian, and Susmit Jha. 2025. Calibrating uncertainty quantification of multi-modal llms using grounding. *arXiv preprint arXiv:2505.03788*.
- Trilok Padhi, Ugur Kursuncu, Yaman Kumar, Valerie L Shalin, and Lane Peterson Fronczek. 2024. Enhancing cross-modal contextual congruence for crowdfunding success using knowledge-infused learning. In *2024 IEEE International Conference on Big Data (BigData)*, pages 1844–1853. IEEE.
- 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*. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 2783–2796, Online. Association for Computational Linguistics.
- Shraman Pramanick, Shivam Sharma, Dimitar Dimitrov, Md. Shad Akhtar, Preslav Nakov, and Tanmoy Chakraborty. 2021b. *MOMENTA: A multimodal framework for detecting harmful memes and their targets*. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 4439–4455, Punta Cana, Dominican Republic. 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 Rohini K. Srihari. 2023. Hate speech detection in low resource indo-aryan languages. In *Fire*.
- 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.
- Ursula Kristin Schmid. 2025. Humorous hate speech on social media: A mixed-methods investigation of users' perceptions and processing of hateful memes. *New Media & Society*, 27(3):1588–1606.
- Hendra Setiawan. 2024. Accurate knowledge distillation via n-best reranking. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 1330–1345. Association for Computational Linguistics.
- 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*.
- Emily A Vogels. 2021. *The state of online harassment*, volume 13. Pew Research Center Washington, DC.
- 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 pre-trained 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. 2016. Hateful symbols or hateful people? predictive features for hate speech detection on Twitter. In *Proceedings of the NAACL Student Research Workshop*, pages 88–93, San Diego, California. Association for Computational Linguistics.
- Jingxuan Wei, Linzhuang Sun, Yichong Leng, Xu Tan, Bihui Yu, and Ruifeng Guo. 2024. Sentence-level or token-level? a comprehensive study on knowledge distillation. In *Proceedings of IJCAI 2024*, pages 6531–6540. IJCAI.
- 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.
- Michihiro Yasunaga, Hongyu Ren, Antoine Bosselut, Percy Liang, and Jure Leskovec. 2021. Qagnn: Reasoning with language models and knowledge graphs for question answering. *arXiv preprint arXiv:2104.06378*.
- 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)*.

Appendix

A Methodology

Graph Attention Networks (GAT)

GAT employs an attention mechanism to dynamically weigh neighbor contributions, computing attention coefficients α_{vu} between nodes, focusing on the most important neighbor interactions for node updates:

$$\alpha_{vu} = \text{softmax} \left(\text{LeakyReLU} \left(\mathbf{a}^\top [W h_v \parallel W h_u] \right) \right) \quad (7)$$

where α_{vu} is the attention coefficient between node v and neighbor u , \mathbf{a} is a learnable attention vector, W is the weight matrix applied to node features, $[W h_v \parallel W h_u]$ denotes the concatenation of transformed features of nodes v and u , and LeakyReLU is the non-linear activation function.

GNN Architecture	AUC	Acc.	F1
GAT	85.76	76.80	74.45
R-GCN	86.11	76.30	76.14

Table 6: Ablation study on the GNN Architecture used to create for the Graph embedding. Results are reported for the seen split of the HatefulMemes Dataset on the best performing KID-VLM variant (RoberTa , Hop 2 , 750 nodes and Gated Fusion mechanism)

Graph Pooling: Once the node representations h_v for all nodes $v \in G_W$ are updated through graph encoding (e.g., R-GCN, GAT), we learn a pooled representation of the graph to generate a single vector representing the entire working graph using Mean Pooling:

$$h_{\text{graph}} = \frac{1}{|V_W|} \sum_{v \in V_W} h_v \quad (8)$$

where h_{graph} is the aggregated graph representation, V_W represents the nodes in the graph, $|V_W|$ is the total number of nodes, and h_v is the feature representation of node v . This pooled representation condenses the graph's information into a single vector.

B Datasets

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 7. Representative examples of memes from the dataset are presented in Figure 4.

Labels	Train	Val	Test Seen	Test Unseen
Non-Hateful (0)	5481	253	510	1250
Hateful (1)	3019	247	490	750
Total	8500	500	1000	2000

Table 7: Train, Test and Validation Seen splits for the Hateful Memes Dataset

HarMeme Dataset

We also conducted a series of experiments on the HarMeme Dataset (Pramanick et al., 2021a), 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 9 and 11, respectively. Figure 5 presents several example memes from the dataset. As all the baselines studied use the 2-class version of the Intensity variable i.e. they combine the Somewhat Harmful and Very Harmful classes into a single Harmful class (Refer Table 8) as the Very Harmful class is under-represented in the dataset (~6 %)

Labels	Train Set	Val Set	Test Set
Not Harmful (0)	1949	116	230
Harmful (1)	1062	61	124
Total	3013	177	354

Table 8: Train, Test and Validation splits for the HarMeme Dataset Intensity Variable for 2 label

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 9: Train, Test and Validation splits for the HarMeme Dataset Intensity Variable for 3 labels (Refer Table 10)

Framework	Target Metrics				Intensity Metrics			
	Acc.	F1	Precision	Recall	Acc.	F1	Precision	Recall
CLIP	57.26	35.67	50.29	39.24	80.23	53.60	51.55	57.06
CLIP*	62.10	36.88	39.86	38.88	77.12	50.59	49.03	52.26
PromptHate	73.73	72.74	75.32	73.73	74.58	74.50	78.97	74.58
HateClipper	75.00	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 10: 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.

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 11: Train, Test and Validation splits for the HarMeme Dataset Target Variable (Refer Table 10)

C Fusion Mechanisms

Let $\mathbf{E}_g \in \mathbb{R}^{d_g}$ and $\mathbf{E}_m \in \mathbb{R}^{d_m}$ 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 Gated Fusion applies a gating mechanism to combine the graph and multimodal embeddings. The fusion is computed as:

$$\mathbf{F}_{\text{gated}} = \sigma(\mathbf{W}_g[\mathbf{E}_g \parallel \mathbf{E}_m]) \odot \mathbf{E}_g + (1 - \sigma(\mathbf{W}_g[\mathbf{E}_g \parallel \mathbf{E}_m])) \odot \mathbf{E}_m \quad (9)$$

where $\mathbf{W}_g \in \mathbb{R}^{(d_g+d_m) \times d_g}$ represents learnable weights, $\sigma(\cdot)$ is the sigmoid activation function, \odot denotes element-wise multiplication, and \parallel represents concatenation of the embeddings.

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

$$\mathbf{F}_{\text{bilinear}} = \mathbf{E}_g^T \mathbf{W}_b \mathbf{E}_m \quad (10)$$

where $\mathbf{W}_b \in \mathbb{R}^{d_g \times d_m}$ is a bilinear transformation matrix that models interactions between the modalities.

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

$$\mathbf{F}_{\text{HAN}} = \sum_{l=1}^L \alpha_l (\mathbf{W}_l [\mathbf{E}_g \parallel \mathbf{E}_m]) \quad (11)$$

where α_l is the attention weight at level l , $\mathbf{W}_l \in \mathbb{R}^{(d_g+d_m) \times d}$ is the learnable weight matrix, and L is the total number of hierarchical levels.

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

$$\mathbf{F}_{\text{mult}} = \tanh(\mathbf{W}_m \mathbf{E}_g) \odot \tanh(\mathbf{W}_m \mathbf{E}_m) \quad (12)$$

where $\tanh(\cdot)$ is the hyperbolic tangent activation function, and $\mathbf{W}_m \in \mathbb{R}^{d_g \times d_m}$ is a weight matrix that transforms the embeddings before interaction.

D 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.

E Failure Cases

Our model utilizes caption embeddings for knowledge distillation, which has led to a few misclassifications due to subtle nuances in image-text relationships. Below, we highlight three cases where the model's predictions differed from the ground truth.

E.1 Case 1: Misclassification Due to Unexpected Context

The image in Figure 8 shows a woman with curly hair, wearing a light-colored collared top, and re-

acting with surprise. The caption reads: "*i hear a funny joke about dish washers.*"

The model incorrectly classified this image as toxic. This misclassification likely occurred because jokes about dishwashers are uncommon, and in some cases, household appliances—especially dishwashers—have been used in derogatory jokes about women. Since the meme presents a reaction of surprise to such a joke, the model may have inferred that the joke carried a sexist connotation, even though there is no direct implication of that in the image or text. This suggests the model was influenced by prior associations rather than the actual content of the meme.

E.2 Case 2: Failure to Recognize the Referenced Individual

Figure 9 is a split-image meme. On the top, a middle-aged Black man in a dark suit has a serious expression. On the bottom, a noose made of blue rope is shown against a black background. The caption reads: "*it's what everyone's been waiting for.*"

The model incorrectly classified this image as non-toxic. The likely reason is that the caption did not explicitly mention the man's identity (Barack Obama), which prevented the model from associating the image with its full context. Without recognizing the specific political or racial implications of the noose imagery, the model misinterpreted the meme's intent. This highlights a challenge in entity recognition—when crucial details are missing from the caption, the model struggles to infer context accurately.

E.3 Case 3: Confusion Due to Facial Features and Expression

The meme in Figure 10 shows a woman with a headband and red nail polish, pressing her hands against her face with a distressed expression. The caption reads: "*don't think this cream is working... maybe this will work??... nothing's happening.*"

The model incorrectly classified this image as toxic. This may have been due to the woman's facial features—specifically, her relatively thin eyes while being in distress—which led the model to incorrectly associate the image with offensive stereotypes. This suggests that certain facial expressions and features in training data may have been disproportionately linked to harmful content, causing unintended biases in classification. However, the

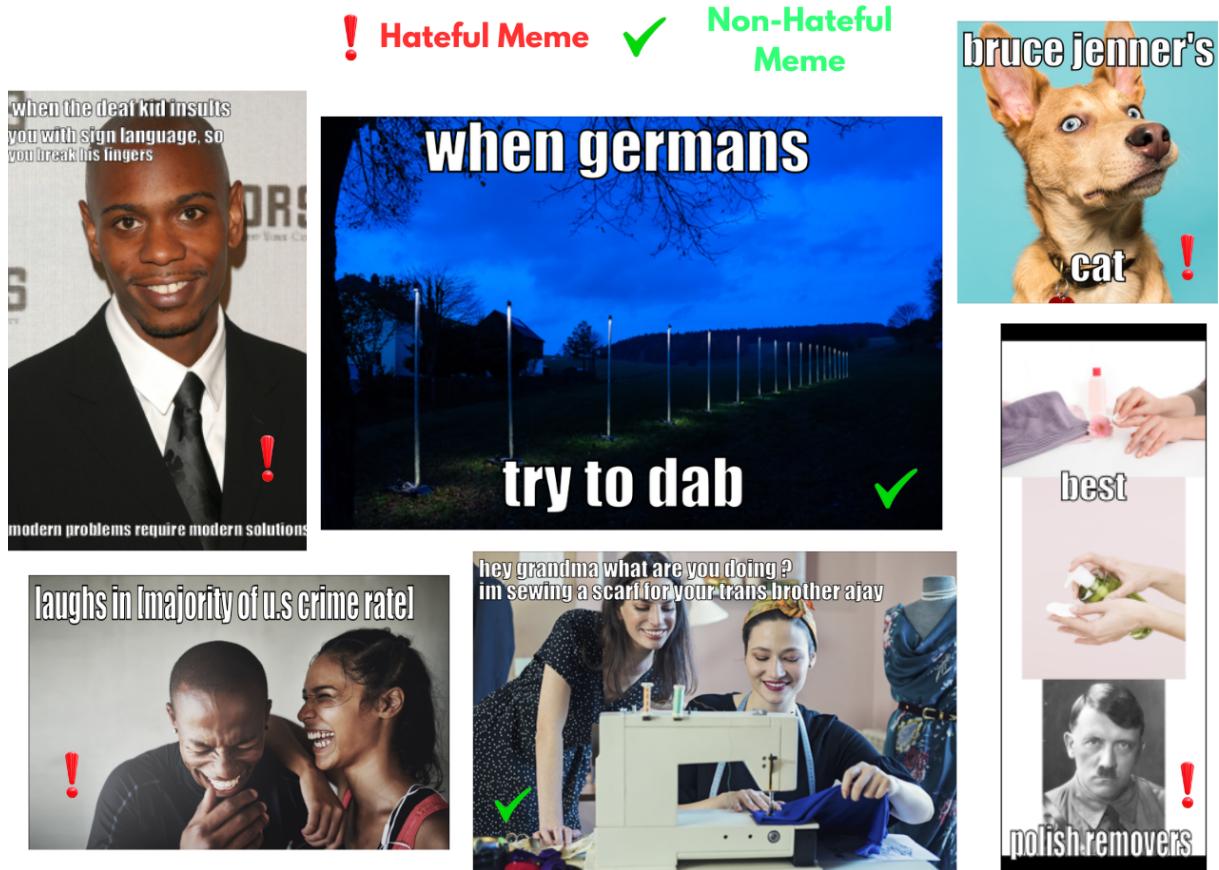


Figure 4: Examples from the Hateful Memes Dataset



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)

image itself is entirely benign, merely depicting frustration with a skincare product.

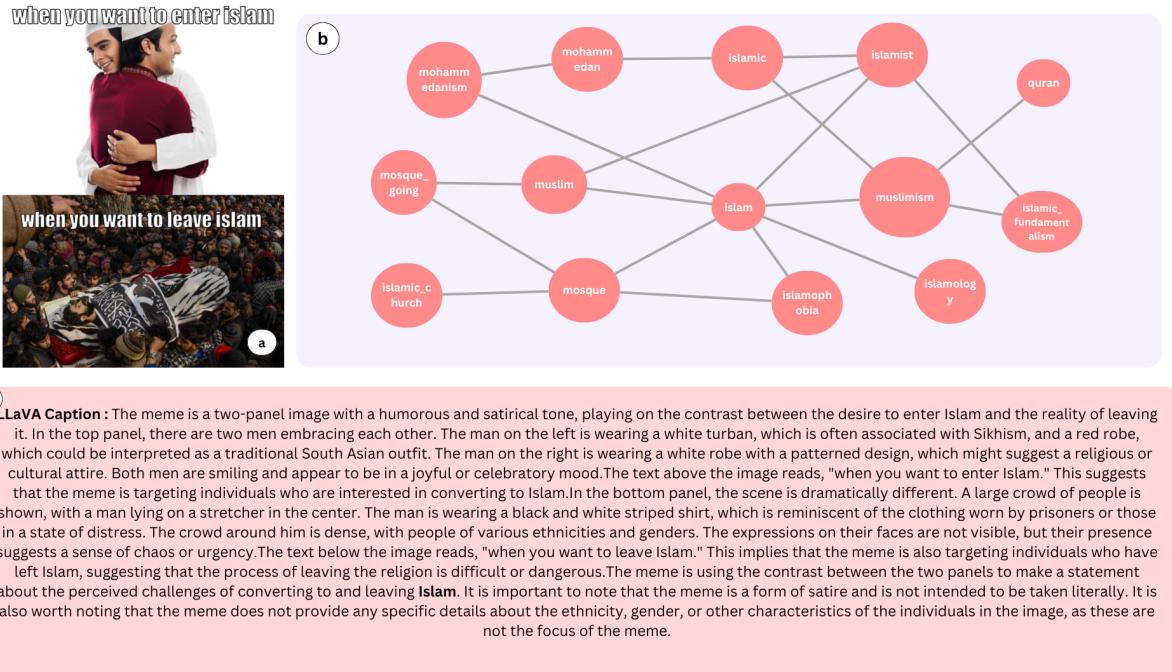


Figure 6: ① Example meme image from the Hateful Meme Dataset. ⑥ Portion of extracted subgraph after Relevancy Scoring and Pruning. ③ Generated LLaVA caption for the meme

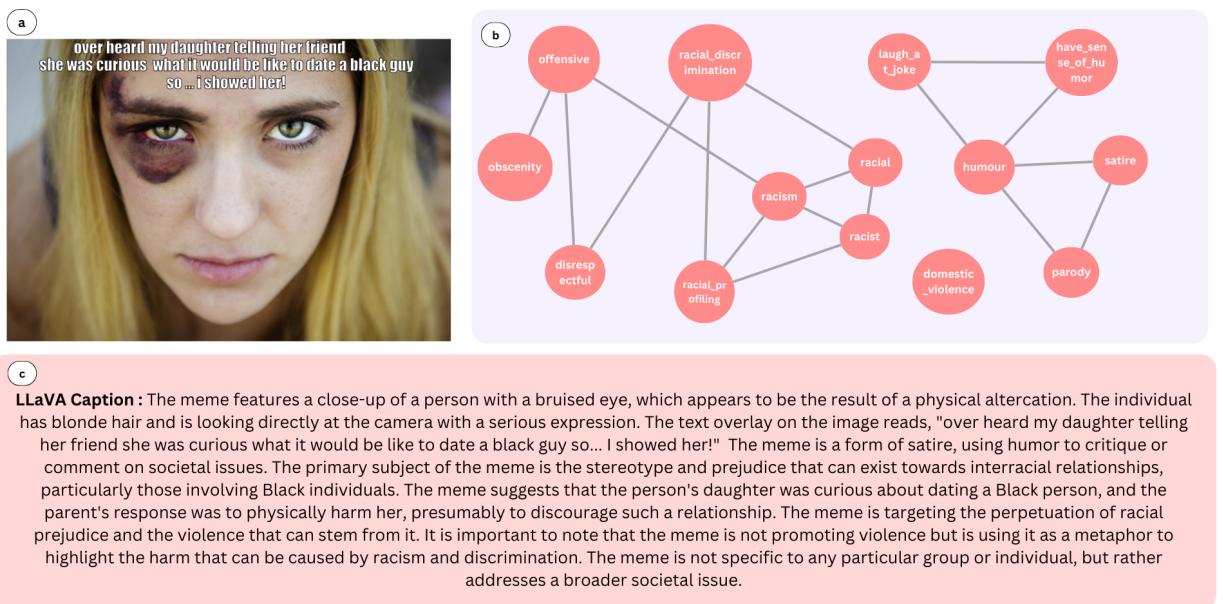


Figure 7: ① Example meme image from the Hateful Meme Dataset. ⑥ Portion of extracted subgraph after Relevancy Scoring and Pruning. ③ Generated LLaVA caption for the meme

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



Figure 8: Meme with surprised woman.

LLaVA caption: The meme features a woman with curly hair, who appears to be in a state of surprise or shock. She is wearing a light-colored top with a collar, and her mouth is open as if she is in the middle of speaking or reacting to something. The background is a neutral, muted gray, which puts the focus on her expression. The text on the meme reads, "*i hear a funny joke about dishwashers*." The humor in the meme seems to be derived from the juxtaposition of the woman's surprised expression with the mundane and unexpected subject of the joke. The meme is likely intended to be humorous by playing on the idea that a joke about dishwashers is not a typical topic for a joke, and the woman's reaction suggests that she is surprised by the subject matter. The meme does not target any specific group or individual, and there are no visible characteristics that would indicate the ethnicity or gender of the woman. The meme is focused on the reaction to the joke, rather than the identity of the person reacting. The primary subject or target of the meme is the unexpected and possibly absurd nature of the joke itself.



Figure 9: Meme with noose juxtaposition.

LLaVA caption: The meme features a split image with two distinct parts. On the left side, there is a photograph of a man with a serious expression. He appears to be middle-aged, with dark skin and short hair. The man is wearing a dark suit and has a slight frown on his face. On the right side of the meme, there is a graphic illustration of a noose with a blue rope. The noose is hanging against a black background, and the phrase "*it's what everyone's been waiting for*" is written above it. The meme seems to be a form of satire or dark humor, using the juxtaposition of the man's serious expression with the noose to create a shocking or ironic effect. The phrase "*it's what everyone's been waiting for*" suggests that the noose represents a long-awaited or anticipated event, which is likely meant to be a commentary on a specific political or social issue. The primary subject or target of the meme is not explicitly stated, but it appears to be directed towards a group of people who are perceived to be waiting for or advocating for a particular outcome. The meme's use of a noose as a symbol may be intended to provoke a strong emotional response, possibly to criticize or satirize the views of the targeted group. It is important to note that the meme contains potentially offensive content, and its use of a noose as a symbol may be insensitive or inappropriate. The meme's message and intent are open to interpretation, and it is not appropriate to use such imagery in a public or professional setting.



Figure 10: Meme with woman applying cream.

LLaVA caption: The meme features a close-up of a person with their eyes closed, hands placed on their face, and a pained or distressed expression. The person appears to be a woman with dark hair, wearing a headband with a patterned design. She has red nail polish on her nails. The background is a plain, light-colored wall. The meme is captioned with a text that reads, "*don't think this cream is working... maybe this will work??... nothing's happening.*" The text is in a casual, conversational font, suggesting a humorous or sarcastic tone. The primary subject of the meme seems to be the process of applying or using a product, such as a cream or lotion, with the implication that the product is not working as expected. The person in the image is likely the target of the meme, as they are shown in a state of frustration or disappointment with the product. The meme does not explicitly identify the ethnicity, gender, or other characteristics of the individual, but it is clear that the person is a woman. The meme does not target any specific group or individual based on their ethnicity, gender, or other characteristics. Instead, it focuses on a relatable experience of product use and the potential for disappointment or frustration when a product does not meet expectations. The meme is likely intended to be humorous or relatable to a wide audience, rather than targeting a specific group.