Uncertainty-Guided Modal Rebalance for Hateful Memes Detection

Chuanpeng Yang[1,2], **Yaxin Liu**[1,2], **Fuqing Zhu**[1,2*], **Jizhong Han**[1], **Songlin Hu**[1,2]
[1]Institute of Information Engineering, Chinese Academy of Sciences
[2] School of Cyber Security, University of Chinese Academy of Sciences
{yangchuanpeng, liuyaxin, zhufuqing, hanjizhong, husonglin}@iie.ac.cn

Abstract

Hateful memes detection is a challenging multimodal understanding task that requires comprehensive learning of vision, language, and cross-modal interactions. Previous research has focused on developing effective fusion strategies for integrating hate information from different modalities. However, these methods excessively rely on cross-modal fusion features, ignoring the modality uncertainty caused by the contribution degree of each modality to hate sentiment and the modality imbalance caused by the dominant modality suppressing the optimization of another modality. To this end, this paper proposes an Uncertainty-guided Modal Rebalance (UMR) framework for hateful memes detection. The uncertainty of each meme is explicitly formulated by designing stochastic representation drawn from a Gaussian distribution for aggregating cross-modal features with unimodal features adaptively. The modality imbalance is alleviated by improving cosine loss from the perspectives of intermodal feature and weight vectors constraints. In this way, the suppressed unimodal representation ability in multimodal models would be unleashed, while the learning of modality contribution would be further promoted. Extensive experimental results demonstrate that the proposed UMR produces the state-of-the-art performance on four widely-used datasets.

Disclaimer: *This paper contains discriminatory content that may be disturbing to some readers.*

1 Introduction

Memes, a form of user-generated content on social media platforms, have become a prevalent way for expressing opinions. Generally, memes consist of an image paired with a humorous caption. However, against a backdrop of current political and socio-cultural fragmentation, a sharply increasing number of individuals are exploiting this format to propagate hate content on platforms by adeptly combining image with text. Therefore, detecting and curbing hateful memes is a particularly urgent research issue.

Figure 1: Examples demonstrate the inherent uncertainty in hateful memes, which is the degree of contribution between modalities to hate sentiment. The left example indicates that identifying hate sentiment should focus on cross-modal features, while the right example should focus on unimodal features.

Previous research on hateful memes detection has employed pre-trained vision-language models for learning vision, language, and cross-modal interactions comprehensively (Das et al., 2020; Muennighoff, 2020; Zhou et al., 2021). Meanwhile, sophisticated fusion techniques (Kiela et al., 2020; Lee et al., 2021; Yang et al., 2023) and external knowledge enhancement methods (Zhu, 2020; Yang et al., 2022; Cao et al., 2022, 2023) have been proposed to further learn the discriminative features of memes. Although the above studies have produced promising progress, they excessively rely on multimodal fusion features, where the inherent uncertainty and imbalance between modalities have not been explicitly considered.

The modality uncertainty is caused by the contribution degree of each modality to hate sentiment. As illustrated in Figure 1, the text in the left meme narrates an incredible story, while the image shows two smiling individuals. The text and image convey completely opposite sentiments. In this case, the

* Corresponding author

Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 4361–4371

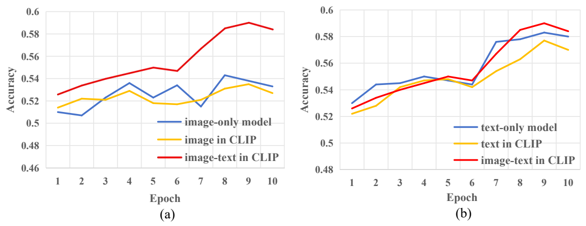

Figure 2: Examples demonstrate the modality imbalance in hateful memes, where text modality suppresses the optimization of image modality. In Figure (a), *image-only model* denotes ViT, *image in CLIP* denotes obtaining image features from CLIP, *image-text in CLIP* denotes obtaining cross-modal features from CLIP. In Figure (b), *text-only model* denotes BERT, *text in CLIP* denotes obtaining text features from CLIP.

multimodal fusion feature can provide additional discriminative information to detect hate sentiment against religion in the meme. Conversely, the text and image in the right meme convey a consistent sentiment, allowing the meme to be identified as non-hateful. However, the learning of the cross-modal fusion features would cause an interaction between *black* in the text and *woman* in the image, leading the model to incorrectly classify it as sexist (Lee et al., 2021). Therefore, quantifying the inherent uncertainty is crucial for determining when unimodal information suffices and when the integration of cross-modal information becomes necessary.

The modality imbalance is caused by the dominant modality suppressing the optimization of another modality. Through experimental analysis, we find that there is a modality imbalance phenomenon between the unimodal features in hateful memes. As shown in Figure 2, the performance of text is closer to multimodal representation compared to image, but the performance of text and image modality in multimodal models is clearly worse than that of the image-only and text-only models, respectively. The above phenomenon indicates that the dominant text modality leads to the suppression of image modality optimization, further resulting in the inability of multimodal models to fully unleash the corresponding discriminative capabilities, thereby affecting the judgment of modality contributions.

To address the above issues, this paper proposes an Uncertainty-guided Modal Rebalance (UMR) framework for hateful memes detection. Specifically, we incorporate a probability distribution to produce a stochastic representation for individual samples, diverging from the deterministic point embeddings employed in current approaches. To simplify modeling, we associate each meme with a Gaussian distribution in a latent space defined by mean and variance parameters. The mean represents the feature, whereas the variance measures the uncertainty. By modeling uncertainty, we flexibly combine discriminative cross-modal and unimodal features. Furthermore, we introduce a cross-modal feature fusion module based exclusively on MLP to capture semantic between images and texts, providing complementary features for hateful memes. Finally, we improve the cosine loss to alleviate the modality imbalance by considering both weight norm and inter-modal constraints. Releasing the unimodal representation ability in multimodal models through modal rebalancing further promotes the learning of modality contributions. The main contributions are summarized as follows:

- We formulate the modality uncertainty and imbalance problems, two critical challenges to hateful memes detection, and present an uncertainty-guided modal rebalance framework to quantify the uncertainty through Gaussian distribution modeling.
- To alleviate the adverse effects of modality imbalance, we improve cosine loss by conducting modality-specific L_2 normalization on both features and weights, fully releasing the representation ability of unimodal in multimodal models to achieve modal rebalancing.
- Extensive experimental results demonstrate that 1) UMR produces the state-of-the-art performance on four widely-used datasets; 2) UMR provides consistent improvement on four vision-language backbones.

2 Related Work

2.1 Hateful Memes Detection

The hateful memes detection task aims to identify detrimental content, including hate, harm, and offense speech. Facebook first proposes the Hateful Memes Challenge (Kiela et al., 2020) to prompt researchers to pinpoint specific categories of hateful content. Prior research has delved into classic dual-stream models that integrate visual and textual features derived from image and text encoders through attention-based mechanisms and various fusion techniques for hate speech classification (Kiela et al., 2020; Das et al., 2020; Lippe et al., 2020; Yang et al., 2023). Recent research has also endeavored to employ data augmentation (Zhou et al., 2021; Zhu, 2020; Lee et al., 2021; Cao et al., 2022, 2023; Yang et al., 2022) and ensemble strategies (Velioglu and Rose, 2020; Sandulescu, 2020) to improve the performance of classifying hateful memes. With the development of hateful memes detection communities, Pramanick *et al.*(Pramanick et al., 2021a) have expanded the categories of hatefulness and introduced two benchmarks pertaining to COVID-19 and US politics. Subsequently, Zhang *et al.*(Zhang et al., 2023) propose TOT to uncover the underlying harm in memes scenario through topology-aware optimal transport. Suryawanshi *et al.*(Suryawanshi et al., 2020) also create a dataset of offensive memes containing abusive messages. Based on this dataset, Lee *et al.*(Lee et al., 2021) propose the DisMultiHate model to disentangle visual and textual representations of memes for understanding. However, the above works overly rely on cross-modal fusion features, where modality uncertainty and imbalance are ignored.

2.2 Uncertainty Learning

The present popular representation learning techniques involve the extraction of features as point representations and aim to position these points as close as possible to the ground truth within a high-level representation space. Nevertheless, there typically exist multiple appropriate point representations, indicating the uncertainty present in representation learning. To tackle this issue, researchers have proposed the use of probability distribution representations to infer diverse solutions and enhance robustness, thereby preventing model overfitting to a single answer. In the domain of natural language processing, Gaussian distribution has been employed to represent words because it effectively captures asymmetric relationships among words (Vilnis and McCallum, 2014). Since then, researchers have explored the use of various distribution families for word representations (Athiwaratkun and Wilson, 2017; Li et al., 2018). In the computer vision domain, to model visual uncertainty, some studies have introduced Gaussian representations into specific tasks, such as person re-identification (Yu et al., 2019), pose estimation (Sun et al., 2020), and face recognition (Chang et al., 2020). More recently, the construction of distributions has yielded progress in generating diverse predictions for cross-modal retrieval in the multimodal field (Chun et al., 2021). However, the uncertainty modeling in the hateful memes detection community remains blank. We are the first to attempt to define the inherent uncertainty between modalities and model each meme as a Gaussian distribution. Furthermore, we consider the issue of modal imbalance and promote the uncertainty learning of memes through modal rebalancing, thereby enhancing the diversity and robustness of the hate detection process.

3 Methodology

3.1 Cross-Modal Feature Encoder

The proposed UMR is illustrated in Figure 3. Taking CLIP (Radford et al., 2021) as an example, for a given image-text pair, we use Vision Transformer and Text Transformer to encode them respectively, and then map them to the same dimension. The encoded image feature is represented as $I \in \mathbb{R}^{l\times d}$, and the text feature is represented as $T \in \mathbb{R}^{k\times d}$.

3.2 Modal Rebalance Module

As discussed in Figure 2, the inconsistency in performance between modalities demonstrates the imbalance, where the modality with worse performance is particularly suppressed. Recently, cosine loss (Ranjan et al., 2017) has been proven effective in reducing intra-class imbalance by L_2 normalization or maximizing cosine similarity scores on features in multimodal fine-grained tasks (Liu et al., 2017; Deng et al., 2019; Xu et al., 2023). Inspired by this prior work, we extend it to the hateful community to alleviate the modality imbalance in hateful memes, focusing on weight norm and inter-modal constraints.

Specifically, we first concatenate the features from the image encoder and the text encoder and obtain the logit score of the intermediate process through a fully connected layer. The vanilla softmax loss can be represented as follows:

$$\mathcal{L}_{\text{vani}} = -\frac{1}{N}\sum_{i=1}^{N}\log\frac{e^{W_{y_i}^{\top}[I_i;T_i]+b_{y_i}}}{\sum_{j=1}^{n}e^{W_j^{\top}[I_i;T_i]+b_j}}, \quad (1)$$

where N represents the batch size, $W \in \mathbb{R}^{2d\times n}$ and $b \in \mathbb{R}^{n}$ represent fully connected layer weight and bias, respectively. n represents the hateful class number. We further divide W into two modality-wise module weights $W^I$and W^T. In this way, we can obtain the logit output $f(x_i)_j$ as follows:

$$f\left(x_i\right)_j = W_j^{I^\top} I_i + W_j^{T^\top} T_i, \quad (2)$$

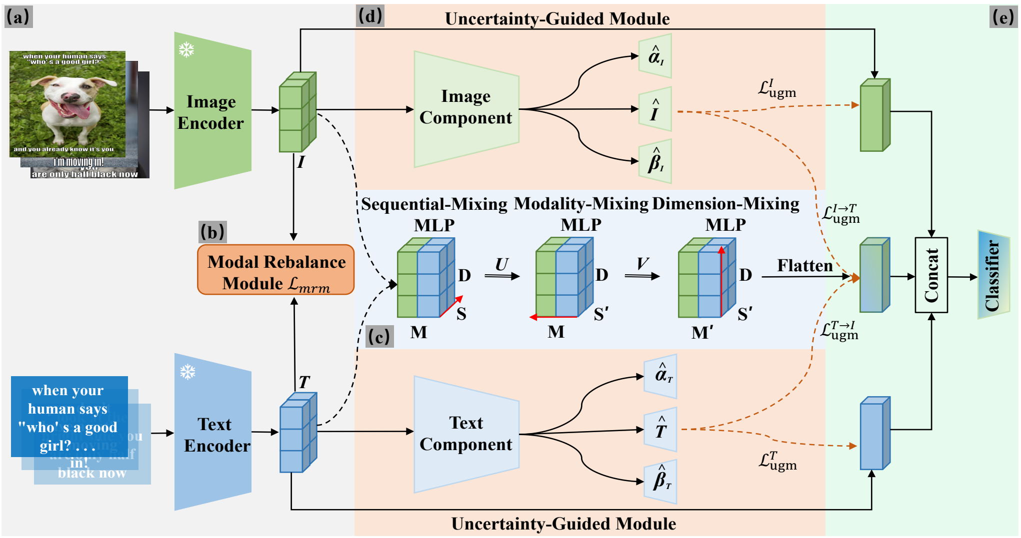

Figure 3: The illustration of the proposed UMR for hateful memes detection. UMR consists of five main components: a) *Cross-Modal Feature Encoder*; b) *Modal Rebalance Module*; c) *Cross-Modal Fusion Module*; d) *Uncertainty-Guided Module*; e) *Hateful Memes Detector*. The cross-modal feature encoder is replaceable and parameter frozen during training.

where $f(x_i)_j$ represents the j-th class of the i-th sample, and bias b is omitted for simplicity. Next, we transform ${W^I_j}^\top I_i + {W^T_j}^\top T_i$ in the logit output to $cos\theta^I_j + cos\theta^T_j$, where $\cos\theta^I_j = \frac{{W^I_j}^\top I_i}{\|W^I_j\|\cdot\|I_i\|}$ and similar for $\cos\theta^T_j$. θ_j is the angle between the weight and the feature. Following previous cosine loss (Wang et al., 2018; Deng et al., 2019; Xu et al., 2023) in fine-grained learning, We fix the modality-wise weights to 1 through L_2 normalization and re-scale the embedding features to s. Normalizing the features and weights ensures that the prediction only depends on the angle between the feature vector and weight vector. Finally, the loss of the modal rebalancing module is defined as:

$$\mathcal{L}_{\text{mrm}} = -\frac{1}{N}\sum_{i=1}^{N}\log\frac{e^{s\cdot\left(\cos\theta^I_{y_i}+\cos\theta^T_{y_i}\right)}}{\sum_{j=1}^{n} e^{s\cdot\left(\cos\theta^I_j+\cos\theta^T_j\right)}}. \tag{3}$$

Through the naive trigonometric transformation, equation 2 can be rewritten as follows:

$$f\left(x_i\right)_j = \cos\theta^I_j + \cos\theta^T_j = 2\cos(\frac{\theta^I_j+\theta^T_j}{2})\cdot\cos(\frac{\theta^I_j-\theta^T_j}{2}). \tag{4}$$

This formula suggests that the logit score will only reach a high value when both modalities exhibit high confidence, meaning θ^I_j and θ^T_j are both small. This acts as a cooperative constraint. Additionally, θ^I_j and θ^T_j must be similar, with their difference remaining small, adhering to the symmetric constraint. This necessitates a more balanced optimization of unimodal features. Therefore, the modal rebalancing module can fully unleash the unimodal representation ability in multimodal models, providing support for subsequent learning of cross-modal features and modality contributions.

3.3 Cross-Modal Fusion Module

Cross-modal fusion has the capability to capture semantic interactions between different modalities, providing complementary features for hate memes detection. This is particularly valuable when image and text feature representations within memes convey conflicting sentiments. Recently, Multi-layer Perceptron (MLP)-based models are proposed for vision tasks. By substituting MLP (*e.g.*, MLP-mixer (Tolstikhin et al., 2021) and ResMLP (Touvron et al., 2022)) for the self-attention mechanism, significantly reducing computational costs while maintaining high performance. However, the above models only contain two independent MLPs, one processing the sequential length and another processing the dimension size. CubeMLP (Sun et al., 2022) is the first to transfer it to multimodal feature processing. It adds an additional MLP module to handle modality features.

Inspired by the above MLP-based models, we naturally extend it to feature fusion of hateful memes. Specifically, we first concatenate the unimodal features to comprise a multimodal tensor $C \in \mathbb{R}^{S\times M\times D}$, where S represents the sequential length, M denotes the number of modalities, and

D indicates the size of the feature dimension. Then, the multimodal features are passed to the stacked three MLP units for mixing. Each MLP unit comprises two fully-connected layers followed by a GELU (Hendrycks and Gimpel, 2016) nonlinear activation to mix the multimodal features along the respective axes. A residual connection is employed in the unit according to (Touvron et al., 2022). Taking the first sequential-mixing MLP as an example, tensor $C \in \mathbb{R}^{S\times M\times D}$ can be viewed as a set of vectors of $C_{*,m,d} \in \mathbb{R}^{S\times 1\times 1}$, where $(m, d) \in \{(1, 1), (1, 2), ..., (2, 1), (2, 2), ..., (M, D)\}$. Here, $C_{*,m,d}$ represents the vector of m-th modality and d-th dimension. Each fully-connected layer in the sequential-mixing MLP unit can be represented as:

$$\mathrm{FC}_S(C_{*,m,d}) = W_S C_{*,m,d} + b_S, \tag{5}$$

where S' is the reduced dimension along the S-axis, which is set as a hyperparameter. All $C_{*,m,d}$ share the parameters W_S and b_S. Therefore, the complete sequential-mixing MLP can be represented as:

$$\begin{aligned} U_{*,m,d} = \mathrm{LayerNorm}(&\mathrm{FC}_S(\mathrm{GELU}(\mathrm{FC}_S(C_{*,m,d}))) \\ &+ C_{*,m,d}), \end{aligned} \tag{6}$$

where the output tensor $U \in \mathbb{R}^{S'\times M\times D}$ can be considered as a set of vectors of $U_{*,m,d} \in \mathbb{R}^{S'\times 1\times 1}$.

Similar to the first MLP unit along the S-axis, the output $V \in \mathbb{R}^{S'\times M'\times D}$ from the second MLP unit along the M-axis can be regarded as a set of vectors $V_{s,*,d} \in \mathbb{R}^{1\times M'\times 1}$. The output $G \in \mathbb{R}^{S'\times M'\times D'}$ from the third MLP unit on the D-axis can be regarded as a set of vectors $G_{s,m,*} \in \mathbb{R}^{1\times 1\times D'}$. Here, M' and D' are the reduced dimensions along the M-axis and D-axis, respectively. The indices (s, d) range over $\{(1, 1), (1, 2), ..., (2, 1), (2, 2), ..., (S', D)\}$ and (s, m) range over $\{(1, 1), (1, 2), ..., (2, 1), (2, 2), ..., (S', M')\}$. Finally, the modality-mixing MLP and the dimension-mixing MLP can be represented as:

$$\begin{aligned} V_{s,*,d} = \mathrm{LayerNorm}(&\mathrm{FC}_M(\mathrm{GELU}(\mathrm{FC}_M(U_{*,m,d}))) \\ &+ U_{*,m,d}), \end{aligned} \tag{7}$$

$$\begin{aligned} G_{s,m,*} = \mathrm{LayerNorm}(&\mathrm{FC}_D(\mathrm{GELU}(\mathrm{FC}_D(V_{s,*,d}))) \\ &+ V_{s,*,d}), \end{aligned} \tag{8}$$

where $G \in \mathbb{R}^{S'\times M'\times D'}$ is the mixed cross-modal feature representation.

3.4 Uncertainty-Guided Module

As shown in Figure 1, hateful memes detection should be aware of the uncertainty between modalities. However, for each given input sample, the unimodal features is deterministic. Therefore, we use the probability distribution $P_{\mathbf{z}|x}$ to capture the uncertainty of input embeddings, using the embeddings $\mathbf{z}$ (representing image I or text T) as estimates for the mean of the desired distribution $P_{\mathbf{z}|x}$. The distribution $P_{\mathbf{z}|x}$ can be represented as a parametric distribution $P_{\mathbf{z}|x}(\mathbf{z}|\hat{\mathbf{z}}, \hat{\theta})$ where the parameters can be estimated (Lakshminarayanan et al., 2017; Upadhyay et al., 2023). Therefore, we introduce two modality-specific components to estimate parameters $\hat{\mathbf{z}}, \hat{\theta}$. To ensure that the mean of the distribution estimated by modality-specific components aligns the point estimates generated by the frozen encoders, we establish a probabilistic reconstruction task for the embeddings within each modality. Specifically, for a given sample x, we extract the embedding $\mathbf{z}$ using the frozen encoder. Subsequently, the modality-specific component learns to reconstruct the $\mathbf{z}$, producing a reconstruction denoted as $\hat{\mathbf{z}}$). The modality-specific component is trained by maximizing the likelihood.

$$\zeta^* = \underset{\zeta}{\mathrm{argmax}} \prod_{i=1}^{N} \frac{\hat{\beta}_i e^{-(|\hat{\mathbf{z}}_i - \mathbf{z}_i|/\hat{\alpha}_i)^{\hat{\beta}_i}}}{2\hat{\alpha}_i \Gamma\left(1/\hat{\beta}_i\right)}, \tag{9}$$

where ζ represents the parameters of the component. Γ represents the Gamma function. $\frac{\hat{\beta}_i e^{-(|\hat{\mathbf{z}}_i - \mathbf{z}_i|/\hat{\alpha}_i)^{\hat{\beta}_i}}}{2\hat{\alpha}_i \Gamma(1/\hat{\beta}_i)}$ represents the Generalized Gaussian Distribution (GCD), denoted as $\mathcal{G}$, which is capable of modeling heavy-tailed distributions. It's worth noting that the Gaussian and Laplace distributions are special cases of $\mathcal{G}$ with parameters $\alpha = 1, \beta = 2$ and $\alpha = 1, \beta = 1$, respectively. The variables $\hat{\mathbf{z}}, \hat{\alpha}, \hat{\beta}$ denote the predicted mean, scale, and shape parameters of $\mathcal{G}$ obtained from modality-specific components for the given input $\mathbf{z}_i$. We determine modality-specific optimal parameters by minimizing negative log-likelihood, equivalent to Equation 9. Given $\mathbf{z}$ and the predicted values $\hat{\mathbf{z}}, \hat{\alpha}, \hat{\beta}$, the loss can be expressed as:

$$\mathcal{L}_{\text{ugm}}(\zeta) = \left(\frac{|\hat{\mathbf{z}} - \mathbf{z}|}{\hat{\alpha}}\right)^{\hat{\beta}} - \log \frac{\hat{\beta}}{\hat{\alpha}} + \log \Gamma\left(\frac{1}{\hat{\beta}}\right), \tag{10}$$

where image-specific component loss is represented as $\mathcal{L}^{I}_{\text{ugm}}(\zeta_I)$, and text-specific component loss is represented as $\mathcal{L}^{T}_{\text{ugm}}(\zeta_T)$.

Moreover, there is a phenomenon in hateful memes where the same image corresponds to different texts and the same text corresponds to different images. To this end, we ensure that the output distributions of image and text embeddings related to similar concepts remain close to each other.

$$\mathcal{L}_{\text{ugm}}^{I\to T}(\zeta_I,\zeta_T) = \left(\frac{|\hat{\mathbf{z}}_I - \mathbf{z}_T|}{\hat{\alpha}_I}\right)^{\hat{\beta}_I} - \log\frac{\hat{\beta}_I}{\hat{\alpha}_I} + \log\Gamma\left(\frac{1}{\hat{\beta}_I}\right), \tag{11}$$

$$\mathcal{L}_{\text{ugm}}^{T\to I}(\zeta_I,\zeta_T) = \left(\frac{|\hat{\mathbf{z}}_T - \mathbf{z}_I|}{\hat{\alpha}_T}\right)^{\hat{\beta}_T} - \log\frac{\hat{\beta}_T}{\hat{\alpha}_T} + \log\Gamma\left(\frac{1}{\hat{\beta}_T}\right). \tag{12}$$

The overall objective of the uncertainty-guided module is designed as:

$$\mathcal{L}_{\text{ugm}}(\zeta_I,\zeta_T) = \mathcal{L}_{\text{ugm}}^{I}(\zeta_I) + \mathcal{L}_{\text{ugm}}^{T}(\zeta_T) + \mathcal{L}_{\text{ugm}}^{I\to T} + \mathcal{L}_{\text{ugm}}^{T\to I}. \tag{13}$$

Finally, the uncertainty of different modalities in the sample x_i can be quantified by estimating the unimodal distribution as follows:

$$\hat{\sigma}_I^2 = \frac{\hat{\alpha}_I^2\Gamma(3/\hat{\beta}_I)}{\Gamma(1/\hat{\beta}_I)}, \tag{14}$$

$$\hat{\sigma}_T^2 = \frac{\hat{\alpha}_T^2\Gamma(3/\hat{\beta}_T)}{\Gamma(1/\hat{\beta}_T)}, \tag{15}$$

$$\lambda_i = \text{Sigmoid}\left(\frac{\hat{\sigma}_I^2 + \hat{\sigma}_T^2}{2}\right), \tag{16}$$

where λ_i represents the uncertainty score. The sigmoid function is employed to map these scores to the range $[0,1]$. The uncertainty score λ_i serves as a weight that governs the fusion of unimodal and cross-modal features. Specifically, the uncertainty-guided module adaptively emphasizes cross-modal features and reduces the influence of unimodal features when uncertainty is high, and does the opposite when uncertainty is low.

3.5 Hateful Memes Detector

We flatten the mixed multimodal features and utilize the uncertainty score λ_i to guide the feature fusion process. Specifically, the cross-modal feature is multiplied by λ_i and each unimodal feature is multiplied by $1-\lambda_i$. The resulting fused feature F_i is then fed into the hateful memes classifier. Cross-entropy loss $\mathcal{L}_{\text{task}}$ is employed for hateful memes detection.

$$F_i = \lambda_i G_i \oplus (1-\lambda_i) I_i \oplus (1-\lambda_i) T_i, \tag{17}$$

$$\hat{y}_i = \text{Softmax}(\text{FC}(F_i)), \tag{18}$$

$$\mathcal{L}_{\text{task}} = -\frac{1}{N}\sum_{i=1}^{N} y_i \log(\hat{y}_i). \tag{19}$$

Table 1: Statistics of Hateful, Harmful-C, Harmful-P and Offensive memes datasets.

Datasets	#Training	#Validation	#Test
Hateful	Hateful(3,050) Non-Hateful(5,450)	Hateful(250) Non-Hateful(250)	Hateful(500) Non-Hateful(500)
Harmful-C	Harmful(1064) Non-Harmful(1949)	Harmful(61) Non-Harmful(116)	Harmful(124) Non-Harmful(230)
Harmful-P	Harmful(1486) Non-Harmful(1534)	Harmful(86) Non-Harmful(91)	Harmful(173) Non-Harmful(182)
Offensive	Offensive(187) Non-Offensive(258)	Offensive(58) Non-Offensive(91)	Offensive(58) Non-Offensive(91)

Finally, we combine the aforementioned modules to optimize the overall objective function of UMR framework:

$$\mathcal{L}_{\text{Loss}} = \mathcal{L}_{\text{task}} + \mathcal{L}_{\text{mrm}} + \mathcal{L}_{\text{ugm}}. \tag{20}$$

4 Experiments

4.1 Datasets

The experiment is conducted on four publicly available datasets as follow: **Hateful memes** which was created as part of the Hateful Memes Challenge 2020 for multimodal hateful detection and published in (Kiela et al., 2020), containing $10K$ memes with binary labels. **Harmful-C memes** and **Harmful-P memes** are respectively related to COVID-19 and United States politics, and published in (Pramanick et al., 2021a) and (Pramanick et al., 2021b) for multimodal harmful detection, each containing nearly $3.5K$ memes with binary labels. **Offensive memes** is related to the 2016 United States presidential election and published in (Suryawanshi et al., 2020) for multimodal offensive detection, containing nearly $1K$ memes with binary labels. The statistics are shown in Table 1.

4.2 Implementation Details

In the cross-modal feature encoder, we use four multimodal backbone models to initialize the image and text encoders, including CLIP ViT-L/14 (Radford et al., 2021), ALBEF ViT-B/16 (Li et al., 2021), BLIP ViT-B/16 (Li et al., 2022) and BLIP-2 ViT-L FlanT5$_{\text{XL}}$ (Li et al., 2023). In the cross-modal fusion module, we set S to 100, which involves zero padding shorter sequences and truncating longer sequences for sequence size matching. The value of M is fixed at 2, as we only have two involved modalities. The dimension D is set to 256, consistent with the output dimension of the projection layer. Additionally, The S', M' and D' are set

Table 2: The performance comparison on Hateful memes. Red represents the best performance, and blue represents the suboptimal performance.

Models	Acc. ↑	AUROC ↑
ViT (Dosovitskiy et al., 2020)	54.30	60.74
BERT (Devlin et al., 2019)	58.30	65.20
MMBT-Region (Kiela et al., 2019)	67.66	73.82
ViLBERT (Lu et al., 2019)	65.27	73.32
Visual BERT (Li et al., 2019)	66.67	74.42
DisMultiHate (Lee et al., 2021)	71.26	79.89
PromptHate (Cao et al., 2022)	72.98	81.45
CDKT (Yang et al., 2022)	76.50	83.74
CLIP (Radford et al., 2021)	59.00	68.30
ALBEF (Li et al., 2021)	68.30	80.79
BLIP (Li et al., 2022)	68.80	74.93
BLIP-2 (Li et al., 2023)	59.70	64.72
UMR_{CLIP}	66.60	78.98
$\text{UMR}_{\text{ALBEF}}$	77.20	85.64
UMR_{BLIP}	74.30	82.37
$\text{UMR}_{\text{BLIP-2}}$	67.40	76.94

Table 3: The performance comparison on Harmful-C and Harmful-P memes.

Models	Harmful-C			Harmful-P		
	Acc. ↑	F1 ↑	MMAE ↓	Acc. ↑	F1 ↑	MMAE ↓
ViT (Dosovitskiy et al., 2020)	68.73	67.81	0.2648	71.19	70.73	0.2481
BERT (Devlin et al., 2019)	70.06	69.92	0.2573	77.97	77.92	0.2090
ViLBERT (Lu et al., 2019)	78.53	78.06	0.1881	87.25	86.03	0.1276
Visual BERT (Li et al., 2019)	81.36	80.13	0.1857	86.80	86.07	0.1318
MOMENTA (Pramanick et al., 2021b)	83.82	82.80	0.1743	89.84	88.26	0.1314
TOT (Zhang et al., 2023)	87.01	85.93	0.1634	91.55	91.29	0.1245
CLIP (Radford et al., 2021)	73.45	72.61	0.2508	83.02	82.83	0.1604
ALBEF (Li et al., 2021)	78.75	77.67	0.1944	87.86	87.04	0.1330
BLIP (Li et al., 2022)	82.77	80.93	0.1774	89.45	88.19	0.1297
BLIP-2 (Li et al., 2023)	84.75	84.01	0.1397	87.04	87.04	0.1292
UMR_{CLIP}	79.10	77.91	0.1943	88.73	88.72	0.1113
$\text{UMR}_{\text{ALBEF}}$	83.62	82.59	0.1614	90.99	90.98	0.0898
UMR_{BLIP}	87.85	86.99	0.1195	92.11	92.11	0.0786
$\text{UMR}_{\text{BLIP-2}}$	86.76	86.66	0.1339	90.42	90.40	0.0963

Table 4: The performance comparison on Offensive memes.

Models	F1 ↑	Pre. ↑	Rec. ↑
ViT (Dosovitskiy et al., 2020)	46.87	45.45	48.39
BERT (Devlin et al., 2019)	52.17	47.37	58.06
StackedLSTM+VGG16 (Suryawanshi et al., 2020)	46.30	37.30	61.10
BiLSTM+VGG16 (Suryawanshi et al., 2020)	48.00	48.60	58.40
CNNText+VGG16 (Suryawanshi et al., 2020)	46.30	37.30	61.10
ERNIE-VIL (Yu et al., 2021)	53.10	54.30	63.70
DisMultiHate (Lee et al., 2021)	64.60	64.50	65.10
CLIP (Radford et al., 2021)	58.94	60.98	59.07
ALBEF (Li et al., 2021)	59.00	59.71	58.91
BLIP (Li et al., 2022)	60.46	62.69	60.49
BLIP-2 (Li et al., 2023)	65.09	68.32	68.13
UMR_{CLIP}	63.28	63.68	64.38
$\text{UMR}_{\text{ALBEF}}$	66.50	66.70	66.36
UMR_{BLIP}	67.12	67.00	67.29
$\text{UMR}_{\text{BLIP-2}}$	69.96	70.30	69.73

to 10, 2 and 32, respectively. In the uncertainty-guided module, both the image and text components consist of Multi-Layer Perceptrons. Each MLP comprises an input layer, transforming the embedding dimension to 256, a hidden layer of size 256, and an output layer, converting from 256 back to the embedding dimensions. For the hateful memes detector, the intermediate feature dimension of the detector is 64 and the dropout rate is set to 0.4. For the above backbone models, the initial learning rate is set to 1e-5, 3e-5, 2e-5 and 1e-4, respectively. The size of the minibatch is set to 16. Each dataset is trained for 10 epochs. The UMR framework is trained on a single A800 GPU.

4.3 Evaluation Metrics

For the evaluation of Hateful memes, we adopt the methodology outlined in (Kiela et al., 2020), employing the Area Under the Receiver Operating Characteristic curve (AUROC) and accuracy (Acc.) as evaluation metrics. The AUROC serves as the primary metric. For Harmful-C and Harmful-P memes, we adhere to the evaluation protocol established by (Pramanick et al., 2021b). Here, we utilize (Acc.), Macro-F1 (F1), and Macro-Averaged Mean Absolute Error (MMAE) as evaluation metrics. In the case of Offensive memes, we follow the evaluation procedure described in (Suryawanshi et al., 2020), employing F1 score, precision (Pre.), and recall (Rec.) as evaluation metrics.

4.4 Experimental Results

Comparison with the baselines. To assess the efficacy of the proposed UMR framework, we utilize four vision-language models (CLIP, ALBEF, BLIP, and BLIP-2) as the backbone of UMR, which also serve as the baselines in this study. As depicted in Table 2 to Table 4, it is evident that UMR demonstrates a notable improvement over the respective baselines. All parameters of the baselines are fine-tuned except for BLIP-2, so the trainable parameters of UMR are fewer compared to the corresponding baseline due to the encoder being frozen. Specifically, for hateful memes, AUROC shows an increase of $+10.68\%$, $+4.85\%$, $+7.44\%$ and $+12.22\%$ on each backbone. For harmful-C memes, MMAE is improved by 0.0565, 0.0330, 0.0579 and 0.0058, respectively. For harmful-P memes, MMAE is improved by 0.0491, 0.0432, 0.0511 and 0.0329, respectively. For offensive memes, F1 is increased by $+5.73\%$, $+7.50\%$, $+6.66\%$ and $+4.87\%$, respectively. The stable improvement demonstrates the effectiveness of modeling modality uncertainty and imbalance. Moreover, the experimental outcomes across multiple backbones highlight the flexible scalability of UMR.

Comparison with the state-of-the-art methods. As this paper simultaneously evaluates four datasets for hateful memes detection, the compar-

Table 5: Ablation study evaluated on the Hateful, Harmful-C, Harmful-P and Offensive memes.

Models	**Hateful** (ALBEF)		**Harmful-C** (BLIP)		**Harmful-P** (BLIP)		**Offensive** (BLIP-2)		
	Acc. ↑	**AUROC ↑**	**F1 ↑**	**MMAE ↓**	**F1 ↑**	**MMAE ↓**	**F1 ↑**	**Pre. ↑**	**Rec. ↑**
UMR	**77.20**	**85.64**	**86.99**	**0.1195**	**92.11**	**0.0786**	**69.96**	**70.30**	**69.73**
UMR w/o mrm	75.40	83.48	84.45	0.1298	90.41	0.0958	67.36	67.72	68.62
UMR w/o ugm	73.60	81.82	83.39	0.1605	89.01	0.1090	66.38	67.13	66.39
UMR w/o cfm	76.40	84.37	86.80	0.1236	90.68	0.0927	67.91	68.76	68.43

Table 6: Performance comparison of various uncertainty learning methods. UMR-COS and UMR-DIS represent *Cosine* and *Euclidean* distances as the uncertainty metrics, respectively.

Models	**Hateful** (ALBEF)		**Harmful-C** (BLIP)		
	Acc. ↑	**AUROC ↑**	**Acc. ↑**	**F1 ↑**	**MMAE ↓**
UMR	**77.20**	**85.64**	**87.85**	**86.99**	**0.1195**
UMR-COS	75.60	83.55	85.93	85.22	0.1267
UMR-DIS	75.10	82.93	84.98	84.94	0.1281

ison methods used for each dataset vary. These methods are outlined below: For hateful memes, CDKT (Yang et al., 2022) is employed, which is a cross-domain knowledge transfer model. It leverages sarcasm domain knowledge to provide additional discriminative information for the relatively small attack samples. For harmful memes, TOT (Zhang et al., 2023) is utilized. TOT deciphers implicit harm in memes scenarios using topology-aware optimal transport. For offensive memes, DisMultiHate (Lee et al., 2021) is employed. DisMultiHate disentangles target information from memes to improve offense content classification. Compared to CDKT, the proposed UMR demonstrates higher performance without requiring additional domain data. UMR achieves this by adaptively aggregating unimodal and cross-modal features through estimating uncertainty between modalities, without the need for complex feature representation (such as entities and demographic information) as required by TOT and DisMultiHate.

Furthermore, we observe that BLIP-2 could provide optimal performance on offensive memes, while showing disappointing results on hateful memes. The primary reason for this discrepancy lies in the construction of the datasets. Compared to the offensive dataset, the hateful dataset introduces benign confounding factors (Kiela et al., 2020) to confuse hate memes, which is particularly rare.

4.5 Quantitative Analysis

Effectiveness of each component. To assess the influence of each component in UMR, we conduct a series of ablation studies, as depicted in Table 5. It can be observed: 1) Removing the modal

Table 7: Performance comparison of various cross-modal fusion methods. UMR-CAT denotes concatenating unimodal representations directly. UMR-CNN denotes using a convolutional neural network for fusion.

Models	**Hateful** (ALBEF)		**Harmful-C** (BLIP)		
	Acc. ↑	**AUROC ↑**	**Acc. ↑**	**F1 ↑**	**MMAE ↓**
UMR	**77.20**	**85.64**	**87.85**	**86.99**	**0.1195**
UMR-CAT	74.70	82.11	85.84	85.46	0.1252
UMR-CNN	75.30	83.34	86.69	85.97	0.1246

rebalance module (w/o mrm), the performance decreases greatly, indicating that modality imbalance will weaken the effectiveness of subsequent feature fusion and uncertainty learning; 2) Removing the uncertainty-guided module (w/o ugm), performance decreases the most, demonstrating that considering the contribution of each modality to hate sentiment is the most critical factor for hateful memes detection task; 3) Removing the cross-modal fusion module (w/o cfm) and using the attention mechanism to capture dependencies between modalities, performance has no significant changes, verifying that MLP-based cross-modal fusion could maintain higher performance while reducing computational costs.

Uncertainty-guided analysis. Table 6 shows the performance of various uncertainty measurement methods. It is evident that all UMR variants exhibit superior performance, underscoring the importance of uncertainty learning in hateful memes detection. Notably, UMR outperforms UMR-COS and UMR-DIS. This is primarily because UMR generates a stochastic representation for each sample using a Gaussian distribution, whereas UMR-COS and UMR-DIS rely on fixed unimodal representations to calculate distance, failing to capture the uncertainty of distributions.

Cross-modal fusion analysis. As illustrated in Table 7, the performance degradation of UMR-CAT is evident, suggesting that merely concatenating unimodal features without modeling cross-modal interactions is inadequate for effective multimodal representation. UMR-CNN, on the other hand, tends to capture locally confined semantic interactions due to the limited size of the convolution kernel. In contrast, UMR can explore these interactions more globally, leading to improved performance.

4.6 Qualitative Analysis

The purpose of UMR is to model modality uncertainty and alleviate modality imbalance for hateful memes detection. To further understand UMR in-

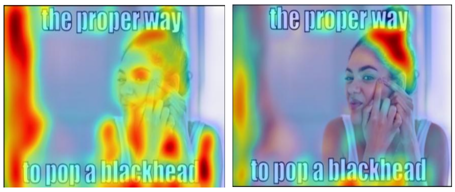

Figure 4: Case analysis of modality uncertainty on Hateful memes.

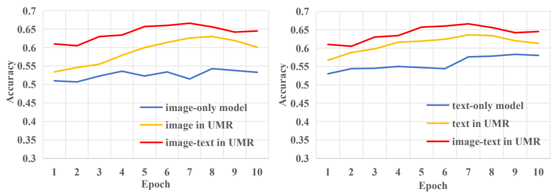

Figure 5: Case analysis of modality rebalance on Hateful memes.

tuitively, we show some cases in Figure 4-5.

Modality uncertainty analysis. We refer to the visualization method in ALBEF and adopt Grad-CAM (Selvaraju et al., 2017) to visualize the third layer cross-attention map of the multimodal encoder in ALBEF. As shown in Figure 4, the left image is ALBEF and the right image is UMR. The keyword is *black*, and the darker the color, the higher the attention of *black* to the image area. The visualization experiment results show that in ALBEF, *black* and the black women in the image, including the specular reflection, are highly focused, leading to misclassification. In UMR, *black* focuses more on the black parts in the image, making the model distinguish it as non-hateful sample.

Modality rebalance analysis. As shown in Figure 5, the performance of text and image in UMR is relatively stable and comparable. This indicates that UMR greatly improves modality imbalance in hateful memes. In addition, the performance of text and image modality in multimodal models has significantly improved compared to image-only and text-only models, respectively.

5 Conclusion

In this paper, a hateful memes detection framework (UMR) is proposed to address the challenges of modality uncertainty and modality imbalance. By extracting stochastic embeddings from a Gaussian distribution to quantify uncertainty, the framework adaptively aggregates cross-modal and unimodal features. By improving cosine loss from weight norm and inter-modal constraints, modality imbalance can be alleviated. Moreover, UMR demonstrates scalability by accommodating various multimodal models as backbones. Experimental results on four widely-used datasets reveal that UMR consistently outperforms baselines and achieves competitive performance compared to existing methods for hateful memes detection. Quantitative analysis further validates the rationality of each component.

Limitations

We would like to highlight some limitations of the proposed method and suggest potential future directions. Firstly, as reported in the experimental results, our method decreases the robustness of the model due to inconsistent backbone performance across different datasets. Secondly, although we have shown the effectiveness of UMR through case studies in this paper, a more comprehensive analysis is required. For example, future work can explore more advanced and interpretable backbone models to enhance the interpretability of hate speech.

References

Ben Athiwaratkun and Andrew Wilson. 2017. Multimodal word distributions. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1645–1656.

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*, MM '23, page 5244–5252.

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

Jie Chang, Zhonghao Lan, Changmao Cheng, and Yichen Wei. 2020. Data uncertainty learning in face recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5710–5719.

Sanghyuk Chun, Seong Joon Oh, Rafael Sampaio De Rezende, Yannis Kalantidis, and Diane Larlus. 2021. Probabilistic embeddings for cross-modal retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8415–8424.

Abhishek Das, Japsimar Singh Wahi, and Siyao Li. 2020. Detecting hate speech in multi-modal memes. *arXiv preprint arXiv:2012.14891*.

Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. 2019. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4690–4699.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*.

Dan Hendrycks and Kevin Gimpel. 2016. Gaussian error linear units (gelus). *arXiv preprint arXiv:1606.08415*.

Douwe Kiela, Suvrat Bhooshan, Hamed Firooz, and Davide Testuggine. 2019. Supervised multimodal bitransformers for classifying images and text. *arXiv preprint arXiv:1909.02950*.

Douwe Kiela, Hamed Firooz, Aravind Mohan, Vedanuj Goswami, Amanpreet Singh, Pratik Ringshia, and Davide Testuggine. 2020. The hateful memes challenge: detecting hate speech in multimodal memes. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, pages 2611–2624.

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. 2017. Simple and scalable predictive uncertainty estimation using deep ensembles. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pages 6405–6416.

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

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International Conference on Machine Learning*.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. 2022. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning*, pages 12888–12900. PMLR.

Junnan Li, Ramprasaath R Selvaraju, Akhilesh Deepak Gotmare, Shafiq Joty, Caiming Xiong, and Steven Hoi. 2021. Align before fuse: Vision and language representation learning with momentum distillation. In *Proceedings of the 34st International Conference on Neural Information Processing Systems*, pages 9694–9705.

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

Xiang Li, Luke Vilnis, Dongxu Zhang, Michael Boratko, and Andrew McCallum. 2018. Smoothing the geometry of probabilistic box embeddings. In *International Conference on Learning Representations*.

Phillip Lippe, Nithin Holla, Shantanu Chandra, Santhosh Rajamanickam, Georgios Antoniou, Ekaterina Shutova, and Helen Yannakoudakis. 2020. A multimodal framework for the detection of hateful memes. *arXiv preprint arXiv:2012.12871*.

Weiyang Liu, Yan-Ming Zhang, Xingguo Li, Zhiding Yu, Bo Dai, Tuo Zhao, and Le Song. 2017. Deep hyperspherical learning. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pages 3953–3963.

Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. Vilbert: pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In *Proceedings of the 33rd International Conference on Neural Information Processing Systems*, pages 13–23.

Niklas Muennighoff. 2020. Vilio: state-of-the-art visio-linguistic models applied to hateful memes. *arXiv preprint arXiv:2012.07788*.

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.

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.

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.

Rajeev Ranjan, Carlos D Castillo, and Rama Chellappa. 2017. L2-constrained softmax loss for discriminative face verification. *arXiv preprint arXiv:1703.09507*.

Vlad Sandulescu. 2020. Detecting hateful memes using a multimodal deep ensemble. *arXiv preprint arXiv:2012.13235*.

Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. 2017. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626.

Hao Sun, Hongyi Wang, Jiaqing Liu, Yen-Wei Chen, and Lanfen Lin. 2022. Cubemlp: An mlp-based model for multimodal sentiment analysis and depression estimation. In *Proceedings of the 30th ACM International Conference on Multimedia*, page 3722–3729.

Jennifer J Sun, Jiaping Zhao, Liang-Chieh Chen, Florian Schroff, Hartwig Adam, and Ting Liu. 2020. View-invariant probabilistic embedding for human pose. In *European Conference on Computer Vision*, pages 53–70.

Shardul Suryawanshi, Bharathi Raja Chakravarthi, Mihael Arcan, and Paul Buitelaar. 2020. Multimodal meme dataset (multioff) for identifying offensive content in image and text. In *Proceedings of the Second Workshop on Trolling, Aggression and Cyberbullying*, pages 32–41.

Ilya Tolstikhin, Neil Houlsby, Alexander Kolesnikov, Lucas Beyer, Xiaohua Zhai, Thomas Unterthiner, Jessica Yung, Andreas Peter Steiner, Daniel Keysers, Jakob Uszkoreit, et al. 2021. Mlp-mixer: An all-mlp architecture for vision. In *Proceedings of Advances in Neural Information Processing systems*, pages 24261–24272.

Hugo Touvron, Piotr Bojanowski, Mathilde Caron, Matthieu Cord, Alaaeldin El-Nouby, Edouard Grave, Gautier Izacard, Armand Joulin, Gabriel Synnaeve, Jakob Verbeek, et al. 2022. Resmlp: Feedforward networks for image classification with data-efficient training. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(4):5314–5321.

Uddeshya Upadhyay, Shyamgopal Karthik, Massimiliano Mancini, and Zeynep Akata. 2023. Probvlm: Probabilistic adapter for frozen vison-language models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 1899–1910.

Riza Velioglu and Jewgeni Rose. 2020. Detecting hate speech in memes using multimodal deep learning approaches: Prize-winning solution to hateful memes challenge. *arXiv preprint arXiv:2012.12975*.

Luke Vilnis and Andrew McCallum. 2014. Word representations via gaussian embedding. *arXiv preprint arXiv:1412.6623*.

Hao Wang, Yitong Wang, Zheng Zhou, Xing Ji, Dihong Gong, Jingchao Zhou, Zhifeng Li, and Wei Liu. 2018. Cosface: Large margin cosine loss for deep face recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5265–5274.

Ruize Xu, Ruoxuan Feng, Shi-Xiong Zhang, and Di Hu. 2023. Mmcosine: Multi-modal cosine loss towards balanced audio-visual fine-grained learning. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

Chuanpeng Yang, Fuqing Zhu, Jizhong Han, and Songlin Hu. 2023. Invariant meets specific: A scalable harmful memes detection framework. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 4788–4797.

Chuanpeng Yang, Fuqing Zhu, Guihua Liu, Jizhong Han, and Songlin Hu. 2022. Multimodal hate speech detection via cross-domain knowledge transfer. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 4505–4514.

Fei Yu, Jiji Tang, Weichong Yin, Yu Sun, Hao Tian, Hua Wu, and Haifeng Wang. 2021. Ernie-vil: Knowledge enhanced vision-language representations through scene graphs. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 3208–3216.

Tianyuan Yu, Da Li, Yongxin Yang, Timothy M Hospedales, and Tao Xiang. 2019. Robust person re-identification by modelling feature uncertainty. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 552–561.

Linhao Zhang, Li Jin, Xian Sun, Guangluan Xu, Zequn Zhang, Xiaoyu Li, Nayu Liu, Qing Liu, and Shiyao Yan. 2023. Tot: topology-aware optimal transport for multimodal hate detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 4884–4892.

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

Ron Zhu. 2020. Enhance multimodal transformer with external label and in-domain pretrain: Hateful meme challenge winning solution. *arXiv preprint arXiv:2012.08290*.