TSR@CASE 2025: Low Dimensional Multimodal Fusion Using Multiplicative Fine Tuning Modules

Sushant Kr. Ray[1], **Rafiq Ali**[2], **Abdullah Mohammad**[2], **Ebad Shabbir**[2], **Samar Wazir**[3†]
[1]University of Delhi, [2]Delhi Skill and Entrepreneurship University, [3]De Montfort University
{skray1331, rafworkacc, abdullah90t, ebadshabbir22}@gmail.com,
samar.wazir@dmu.ac.uk

Abstract

This study describes our submission to the CASE 2025 shared task on multimodal hate event detection, which focuses on hate detection, hate target identification, stance determination, and humour detection on text embedded images as classification challenges. Our submission contains entries in all of the subtasks. We propose FIMIF, a lightweight and efficient classification model that leverages frozen CLIP encoders. We utilise a feature interaction module that allows the model to exploit multiplicative interactions between features without any manual engineering. Our results demonstrate that the model achieves comparable or superior performance to larger models, despite having a significantly smaller parameter count. The source code and model checkpoints are available at github.com/sushant-k-ray/FIMIF

1 Introduction

The landscape of digital communication has evolved dramatically with the widespread adoption of social media platforms, fundamentally transforming how individuals express opinions and share content. This evolution has brought significant challenges in content moderation, particularly in the detection of hate speech that increasingly manifests in the form of memes, which are images with text embedded in them used to convey a message. The CASE (Challenges and Applications of Automated Extraction of Socio-political Events from Text) series has consistently addressed these challenges, with recent editions expanding from text-only analysis to encompass multimodal content understanding (Thapa et al., 2023, 2024).

Building upon the success of previous CASE workshops, the multimodal hate event detection task at CASE 2025 (Thapa et al., 2025a; Hürriyetoğlu et al., 2025) represents a natural progression toward addressing more complex multimodal hate speech detection scenarios.

In this paper, we introduce FIMIF (Feature Interaction for Multimodal Integration and Fusion), a model conceptually similar to MemeCLIP (Shah et al., 2024). We utilise modified residual units to leverage the capabilities of deep neural networks while keeping the performance stable. We introduce a feature interaction module that automatically learns exponential and multiplicative relationships between features, enabling the model to capture higher-order interactions. While MemeCLIP is designed for general downstream tasks on meme images, our model specifically targets meme classification. Our approach relies on aggressive compression of multimodal embeddings to very low dimensions, followed by a multiplicative module that allows for richer feature interactions. We provide comprehensive experimental evaluation demonstrating the effectiveness of our approach.

2 Related Works

Hate Speech Detection: The task of hate speech detection has progressed from lexicon-based or shallow machine learning approaches (Burnap and Williams, 2015; Waseem and Hovy, 2016; Davidson et al., 2017) to deep learning models (Parihar et al., 2021). The advent of large pre-trained language models brought significant improvements in hate speech detection. BERT (Devlin et al., 2019), RoBERTa (Liu et al., 2019), and DistilBERT (Sanh et al., 2020) introduced contextual embeddings that improved performance on social media hate speech detection. These models have achieved state-of-the-art results on benchmarks such as: HateXplain (Mathew et al., 2021), Offensive Language Identification Dataset (OLID) (Rosenthal et al., 2021), Gab Hate Corpus (Kennedy et al., 2022), and Storm-

[†] Corresponding author.

Proceedings of the 8th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Texts associated with RANLP 2025, pages 123–132, Varna, Bulgaria, Sep 13, 2025.
https://doi.org/10.26615/978-954-452-099-1-015

front dataset (de Gibert et al., 2018). These language models are very efficient and powerful in terms of language understanding.

Multimodal Tasks: As harmful content increasingly appears in multimodal forms like memes, research has shifted toward models that process both text and images. Datasets such as Facebook's Hateful Memes (Kiela et al., 2020) and MMHS150K (Gomez et al., 2020) have been instrumental in advancing this field. Some recent multimodal hate speech detection datasets include Harm-C (Pramanick et al., 2021a), Harm-P (Pramanick et al., 2021b), DisinfoMeme (Qu et al., 2022), and CrisisHateMM (Bhandari et al., 2023). Early multimodal systems use separate encoders (e.g., ResNet (He et al., 2016) for images and BERT for text) and combine features through concatenation or attention. Later models rely on fusion strategies to combine these different representations.

Vision Language Models: Vision-Language models aim to learn joint representations of visual and textual inputs, typically trained on large-scale image-text pairs. These models are broadly divided into two categories: Dual-encoder models, and Fusion models.

Dual-encoder models, such as OpenAI's CLIP (Contrastive Language-Image Pre-training) (Radford et al., 2021) and Google's ALIGN (Jia et al., 2021), encode images and text separately and align their embeddings using contrastive loss.

CLIP, in particular, has gained popularity due to its strong zero-shot performance and generalisation ability. Trained on 400 million internet image-text pairs, it can embed both modalities into a shared semantic space.

Low-Rank Multimodal Fusion: One of the key challenges in multimodal learning is the integration of information from multiple modalities. While tensor based fusion methods offer powerful and expressive means of capturing interactions between modalities, they are often computationally expensive and suffer from a rapid increase in parameters, particularly when modelling higher-order interactions across multiple input sources (Zadeh et al., 2017).

To mitigate these challenges, Low-rank Multimodal Fusion (LMF) (Liu et al., 2018) has emerged as a scalable and efficient paradigm. Rather than modelling the full tensor representation, LMF approximates it using modality specific low rank projections, which are then combined using element-wise operations. This dramatically reduces the parameter count and computational overhead while still retaining cross-modal interactions. LMF scales linearly with the number of modalities, in contrast to the exponential growth in traditional fusion approaches. We adapt a similar principle with the use of additive and multiplicative layers.

Highway And Residual Networks: Highway networks (Srivastava et al., 2015) and Residual networks (He et al., 2016) are widely used to improve training stability and depth in deep learning models. Residual layers mitigate vanishing gradients by adding skip connections, while highway layers introduce trainable gates to control information passage. These ideas motivate our use of lightweight residual projections to preserve essential features without over-fitting.

Multiplicative Modules: The Neural Arithmetic Logic Unit (NALU) (Trask et al., 2018) introduces a mechanism for learning arithmetic operations in neural networks using log-space computations to model multiplicative relationships. Several variants of NALU have been proposed to improve stability and expressiveness in different settings (Schlör et al., 2020; Madsen and Johansen, 2020; Heim et al., 2020). We extend NALU to multimodal classification in a residual framework to maintain flexibility while modelling higher-order relationships.

3 Dataset And Tasks

Shah et al. (2024) released a novel multimodal dataset, PrideMM consisting of text embedded images for classification of various aspects of hate against marginalised LGBTQ+ movement, and community in online discourse through images, particularly memes. The dataset is divided into four classification tasks: hate detection, hate target identification, stance determination, and humour detection.

The multimodal hate task at CASE 2025 utilises the PrideMM dataset, focusing on discrimination and hate against the LGBTQ+ community. The dataset is divided into an 80/10/10 train-validation-test split. This is different from the PrideMM dataset, where the split is 85/5/10. OCR of the images is provided as supplementary material to aid in the process of classification.

The following table shows the distribution of the training samples:

Task	Label	Samples	%
Hate	No Hate	2065	50.99%
	Hate	1985	49.01%
Target	Undirected	617	31.08%
	Individual	199	10.03%
	Community	931	46.90%
	Organization	238	11.99%
Stance	Neutral	1166	28.79%
	Support	1527	37.70%
	Oppose	1357	33.51%
Humour	No Humour	1313	32.42%
	Humour	2737	67.58%

Table 1: Distribution of the training samples in the shared task dataset.

3.1 Tasks

The PrideMM dataset focuses on following four subtasks:

Subtask A: Hate Detection. This task aims to identify instances of hate speech in the images. This task focuses on identifying whether the images intentionally convey hateful sentiments. The training data is balanced (1.04 : 1), and contains a total of 4050 data samples.

Subtask B: Hate Target Identification. This task focuses on identifying the targets of hate in hateful images. There are four categories: Undirected, Individual, Community, and Organization. Images are labeled 'Undirected' when they target abstract topics, societal themes, or ambiguous targets. Hateful images targeting specific people are labeled 'Individual'. The label 'Community' is used for instances of hate in images targeting broader social, ethnic, or cultural groups. Images targeting corporate entities, institutions, or similar organizations are labeled 'Organization'.

The training data is extremely imbalanced (3.1 : 1 : 4.7 : 1.2), and contains data samples for only those images which convey hate. As a consequence, only 1985 data samples are available for training.

Subtask C: Stance Determination. This task aims to determine the stance that the image is trying to convey towards the topic. There are three categories: Support, Oppose, and Neutral. The 'Support' label is given to images that express support for the goals of the movement, agree with efforts to promote equal rights for LGBTQ+ individuals, or promote awareness of the movement. The 'Oppose' label is given to images that express disagreement with the goals of the movement, deny the problems faced by individuals who identify as LGBTQ+, or dismiss the need for equal rights and acceptance. The 'Neutral' label is given to images that are contextually relevant to the movement but exhibit neither support nor opposition towards the movement.

The training data is fairly well balanced (1 : 1.31 : 1.16), and contains a total of 4050 data samples.

Subtask D: Humour Detection. This task aims to detect whether the image showcases any form of humour, sarcasm, or satire related to the LGBTQ+ pride movement regardless of whether it presents a light-hearted or insensitive perspective on serious subjects.

The training data is imbalanced (1 : 2.08), and contains a total of 4050 data samples.

4 Methodology

In this section, we describe FIMIF (Feature Interaction for Multimodal Integration and Fusion), our proposed model for meme classification. We utilise the CLIP vision-language model to extract multimodal embeddings that effectively encode the semantic content of memes. Figure 1 illustrates the overall architecture of our model. Below, we describe each component in detail.

Pre-Trained CLIP Model: Similar to MemeCLIP, we leverage CLIP encoders for their strong zero-shot generalisation and effective transfer learning capabilities. The CLIP model consists of an image encoder (E_I) and a text encoder (E_T). We freeze the weights of both encoders to retain the knowledge acquired during pre-training. We utilise CLIP ViT-L/14 image encoder pre-trained on 336x336 images instead of 224x224. 336x336 images can better represent high-frequency information than their 224x224 counterparts. Figure 2 presents an example. Note that the use of 336px variant of CLIP's image encoder does not increase the parameter count of the encoder. The unimodal image and text representations $X_I, X_T \in \mathbb{R}^{768}$ effectively encode the semantic content of a meme and are defined as:

$$X_I = E_I(I);\ X_T = E_T(T) \qquad (1)$$

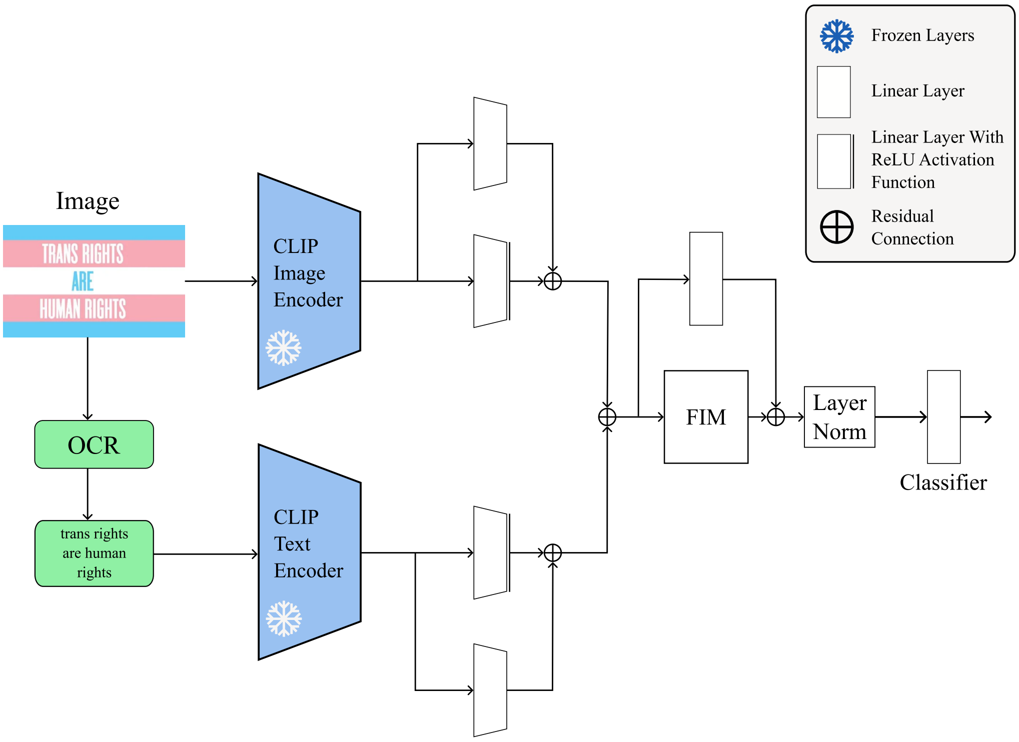

Figure 1: Architecture of our proposed model. Trapeziums are used to represent dimensionality compression.

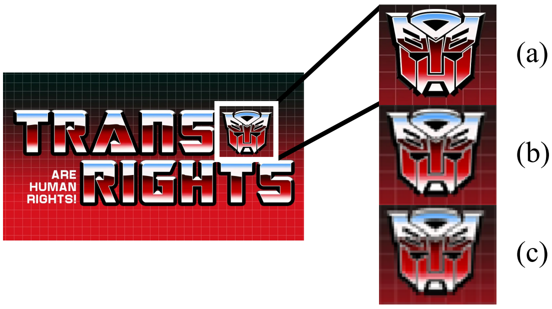

Figure 2: Meme at various resolutions (a) Original resolution (2314x1191) (b) Image downscaled to 336x336 (c) Image downscaled to 224x224. The downscaling and upscaling method used is bicubic interpolation.

where I is the image and T is its corresponding OCR text.

4.1 Linear Residual Projection Layer

Although CLIP is trained to maximise similarity between aligned image-text pairs, the inherently contrastive nature of memes, where visual and linguistic elements often convey conflicting messages, calls for additional adaptation of the embedding space. We hypothesise that only a small subset of elements within the embeddings significantly influence the classification outcome. To capture this, we utilise a modified residual module scheme that effectively compresses the embedding spaces. A regular residual layer is defined as:

$$R(X) = A(X) + X \tag{2}$$

where A is typically a non-linear function. Our modified residual module, although similar to the one described above, performs better in compressing high-dimensional spaces, particularly when combined with lasso regularisation (Tibshirani, 1996). Our residual module is defined as:

$$R(X) = A(X) + B(X) \tag{3}$$

where A is a non-linear function and B is a linear function. The domain and co-domain for both functions are $\mathbb{R}^{768}$ and $\mathbb{R}^{h}$, respectively, where h is a very small number (generally 8, or 16).

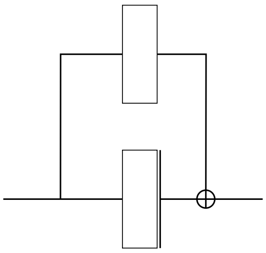

Figure 3: Our Modified Residual Module.

We use a ReLU activation function for A. We utilise this modified residual model for both modalities and combine them to extract linear relations between image and text embeddings. These projection layers result in the bimodal projection $X_{MM} \in \mathbb{R}^h$. Such a layer allows us to leverage the benefits of deep neural network layers while still having the flexibility to use a shallower architecture when required. Our final bimodal residual network is defined as:

$$\begin{aligned} X_{MM} &= R_I(X_I) + R_T(X_T) \\ &= (A_I(X_I) + B_I(X_I)) + \\ &\quad (A_T(X_T) + B_T(X_T)) \end{aligned} \tag{4}$$

4.2 Feature Interaction Module

Since the hidden dimension (h) of the layers is much smaller than CLIP's embedding dimension, we would have difficulty fusing the text and image representations. To capture the non-linear feature interactions in a compact space, we require a multiplicative network. Conceptually, we would like to have a following module:

$$FIM(X) = \begin{bmatrix} M_0(X) \\ M_1(X) \\ \cdots \\ M_{h-1}(X) \end{bmatrix} \tag{5}$$

where,

$$M_i(X) = \prod_{j=0}^{h-1} x_j^{w_{ij}} \tag{6}$$

This module is very generic in nature and can be used for automated feature selection. A module like this, however, would suffer from unstable training due to gradient issues. We design a multiplicative module inspired by Neural Arithmetic Logical Unit (NALU). Rather than directly applying exponentials, we utilise linear arithmetic between inputs in log-space followed by exponentiation. Mathematically, the multiplicative layer can be represented by the following relation:

$$M(X) = B(exp(Wln(ReLU(A(X))+\epsilon))) \tag{7}$$

where A and B are some linear transformation function with both, domain and co-domain, in $\mathbb{R}^h$. Since logarithm of non-positive numbers is undefined, we use ReLU along with some ϵ (10^{-5} in our case). This strictly positive condition, however, prevents us from multiplying a positive and a negative number. While several variants of the NALU, such as iNALU (Schlör et al., 2020) and NAU (Madsen and Johansen, 2020), introduce complex modifications to address this, we propose a simpler alternative that leverages multiple multiplicative layers. We call this the Feature Interaction Module (FIM). Mathematically, it is defined as follows:

$$FIM(X) = M_a(X) \cdot M_b(X) \tag{8}$$

The Feature Interaction Module is shown in the following diagram:

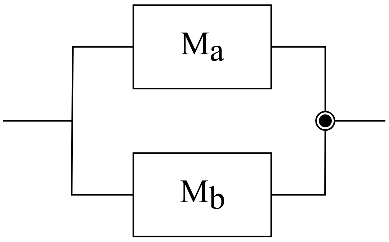

Figure 4: Feature Interaction Module.

We add a residual unit to the FIM in order to allow the network to bypass multiplicative layers if required. The complete residual FIM is defined as:

$$\begin{aligned} F_{MM} &= FIM(X_{MM}) + A(X_{MM}) \\ &= M_a(X_{MM}) \cdot M_b(X_{MM}) + A(X_{MM}) \end{aligned} \tag{9}$$

where A is a linear transformation function with domain and co-domain in $\mathbb{R}^h$.

4.3 Miscellaneous

Classifier: We apply layer normalisation on the outputs of the residual FIM (F_{MM}) before passing it through the classifier. The classifier is a linear transformation function from $\mathbb{R}^h$ to $\mathbb{R}^c$, where c is the number of categories in the given subtask. A softmax function maps the final hidden representations to their respective class probabilities. The predicted class corresponds to the highest probability score.

Class Imbalance: There is a heavy class imbalance in the dataset. To get around this issue, we utilise weighted cross-entropy loss. Further, we utilise minority-class deterministic oversampling for subtask B, where there is an extreme class imbalance. The intuition behind this is to expose the model to more samples from minority classes in order to better classify them. Compared to the high dimensionality of the image and text embeddings

Method	# of trainable	Hate		Target		Stance		Humour	
	Parameters	Accuracy	F1	Accuracy	F1	Accuracy	F1	Accuracy	F1
Gradient Boosting	-	78.90	78.90	59.44	57.39	61.54	60.52	76.13	70.60
FIMIF (submission)	25k - 51k	81.85	81.85	63.05	60.57	62.92	62.91	79.68	76.83
FIMIF (best)	25k - 51k	81.85	81.85	64.66	64.61	64.89	64.32	79.68	76.83

Table 2: Classification performance of different models on shared task dataset across two evaluation metrics: Accuracy, and F1 score. The hidden dimension of the FIMIF model for subtask A is set to 16, while a reduced hidden size of 8 is used for all other subtasks.

Method	# of trainable		Hate			Target			Stance			Humour	
	Parameters	Acc.	AUC	F1	Acc.	AUC	F1	Acc.	AUC	F1	Acc.	AUC	F1
MemeCLIP	2.6M	76.06	**84.52**	75.09	66.12	**81.66**	58.65	62.00	**80.11**	57.98	80.27	**85.59**	77.21
FIMIF (ours)	25k	**78.11**	83.99	**76.43**	**68.42**	75.97	**62.63**	**63.31**	79.84	**59.52**	**80.47**	85.54	**77.54**

Table 3: Classification performance of different models on PrideMM dataset across three evaluation metrics: Accuracy, AUC, and F1 score. Performance metrics for MemeCLIP is sourced from its corresponding paper. The best performance is highlighted in **bold**.

from the CLIP encoders (768 dimensions each), the size of the training set for subtask B is relatively small, consisting of only 1985 samples. This type of high-dimensional data struggles to generalise. Algorithm 1 presents the pseudocode used for upsampling the minority class.

Algorithm 1 Deterministic Class-wise Upsampling.

```
Require: Dataset $D$ of $(x, y)$ pairs, number of classes $C$
Initialise class_samples$[0 \dots C-1] \leftarrow$ empty lists
for all $(x, y) \in D$ do
    Append $(x, y)$ to class_samples$[y]$
end for
$M \leftarrow \max_{c \in \{0,\dots,C-1\}}$ length of class_samples$[c]$
upsampled_dataset $\leftarrow$ empty list
for $c = 0$ to $C-1$ do
    samples $\leftarrow$ class_samples$[c]$
    $n \leftarrow$ length of samples
    $r \leftarrow \lfloor M/n \rfloor$
    for $i = 1$ to $r$ do
        Append all elements of samples to upsampled_dataset
    end for
end for
Shuffle upsampled_dataset
return upsampled_dataset
```

Weight Initialisation: All weights and biases in our model are initialised using the Kaiming-Uniform distribution (He et al., 2015), which helps maintain a stable gradient flow during the initial phases of training. However, for the weight matrix W in eq. 7, we instead use the identity matrix as the initial weights. By using this initialisation, we ensure that the multiplicative interactions introduced by the FIM initially behaves in a linear and interpretable manner. This allows the model to gradually learn multiplicative behaviour only when it is beneficial, rather than being forced into a multiplicative domain from the beginning. This identity initialisation improves the performance and training time by converging in the early training stages. We use the identity and zero matrices as layer-norm initial weights and biases, respectively.

5 Results

We provide results of our model on the test set of the respective subtask in table 2. We use weighted gradient boosting as a baseline for its excellent generalisation capability with high dimensional data.

Method	Acc.	AUC	F1
CLIP	**81.62**	**88.87**	**79.89**
BERT	80.43	88.08	78.90
RoBERTa	77.27	87.95	75.38
DeBERTaV3	79.45	87.94	77.71

Table 4: Classification performance on subtask A (hate) validation set with our model on CLIP's ViT-L/14@336px image encoder and different text encoders.

Along with our submission results, we have also provided the best results we have encountered so far in order to demonstrate the viability of these very low parameter models. Table 3 compares our results on PrideMM dataset against MemeCLIP

Method	Hidden	Hate			Target			Stance			Humour		
	Dim.	Acc.	AUC	F1	Acc.	AUC	F1	Acc.	AUC	F1	Acc.	AUC	F1
CLIP ViT-L/14	8	77.87	86.70	75.82	38.71	64.77	36.27	61.46	79.80	59.64	78.85	84.94	71.87
(Image Only 224x224)	16	79.84	87.40	78.40	55.65	68.67	48.66	58.50	79.33	55.96	**80.04**	84.78	74.82
CLIP ViT-L/14@336px	8	79.05	87.62	77.12	39.92	65.51	37.41	63.44	81.06	60.20	78.46	**85.54**	74.05
(Image Only 336x336)	16	81.03	87.90	79.39	56.85	69.31	49.16	61.46	81.14	59.60	79.64	85.28	**75.85**
CLIP ViT-L/14@336px +	8	**83.20**	88.72	**81.75**	**60.48**	**70.77**	**51.06**	**64.03**	**81.65**	**62.10**	79.45	85.02	74.75
OCR Text	16	81.62	**88.87**	79.89	47.58	66.43	43.27	**64.03**	81.20	61.68	79.64	84.45	74.65

Table 5: Our experiments with the use of CLIP's image encoders on the validation set of shared task dataset. We use three evaluation metrics: Accuracy, AUC, and F1 score. The best performance is highlighted in **bold**.

Method	Hidden	Hate			Target			Stance			Humour		
	Dim.	Acc.	AUC	F1	Acc.	AUC	F1	Acc.	AUC	F1	Acc.	AUC	F1
FIMIF	8	83.20	88.72	81.75	60.48	70.77	51.06	64.03	81.65	62.10	73.72	84.86	70.69
	16	81.62	88.87	79.89	47.58	66.43	43.27	64.03	81.20	61.68	81.23	85.03	75.77
- FIM	8	79.05	89.18	76.98	59.68	70.76	50.28	62.65	81.73	60.45	76.68	85.22	73.34
	16	81.23	89.51	79.59	59.27	70.43	51.52	63.44	81.54	60.89	76.68	85.01	73.03
- Upsampling	8	79.05	89.18	76.98	58.47	71.16	50.84	63.44	81.33	60.91	75.10	85.07	71.17
	16	81.23	89.51	79.59	60.48	71.44	50.21	61.46	81.24	58.72	77.87	85.12	73.08
- Weighted	8	78.66	89.04	76.54	60.48	72.26	51.87	61.66	81.22	57.75	80.63	85.22	73.92
Loss	16	82.02	89.45	80.25	58.87	71.64	49.10	60.08	81.34	57.26	79.64	85.47	74.46

Table 6: Ablation experiments performed on the validation set of given shared task dataset.

Hidden Dim.	Acc.	AUC	F1
4	81.23	88.75	79.24
8	**83.20**	88.72	**81.75**
16	81.62	88.87	79.89
32	77.27	88.86	75.41
64	81.03	**89.22**	79.10
128	80.24	88.31	78.64
256	68.77	89.11	63.14

Table 7: Classification performance on the subtask A (hate) validation set across different hidden dimensions of our model. The best performance is highlighted in **bold**.

on all four subtasks. We compare CLIP's text encoders with other large language models in table 4. These models are trained in a deterministic manner (having no randomness) in order to compare different methods. CLIP's text encoder, despite having a shorter context length of 77 tokens, performs better than BERT, RoBERTa, and DeBERTaV3 (He et al., 2023), each supporting a context length of up to 512 tokens. Table 5 compares the results of our model on CLIP ViT-L/14 224px and 336px image encoders on the validation set of the shared task dataset. Table 7 presents a comparison of our model across different hidden dimensions, showing little to no improvement as the dimension size increases, possibly due to over-fitting. Table 8 reports results on the HarMeme-C dataset (Pramanick et al., 2021a), where our model is compared against several state-of-the-art approaches.

Method	Acc.	F1
MOMENTA (Pramanick et al., 2021b)	83.82	82.80
PromptHate (Cao et al., 2022)	84.47	-
Pro-Cap (Cao et al., 2023)	85.03	-
MemeCLIP (Shah et al., 2024)	84.72	83.74
FIMIF (ours)	**87.01**	**83.94**

Table 8: Performance comparison of meme classification models on the HarMeme-C dataset (binary classification). The best performance is highlighted in **bold**.

5.1 Ablation Study

We have performed our ablation study on the validation set. We compare our model with the one where feature interaction module has been replaced with a linear transformation layer having a non-linear ReLU activation function. The findings in table 6 suggest that the CLIP embeddings of PrideMM dataset is very linear in nature. Due to its residual design, our implementation of feature interaction module is very generic. It can perform just as well, if not better, than a residual module even when the data does not exhibit multiplicative relationships. The difference between these architectures is likely due to the overhead incurred by having a larger number of parameters (3.5 times that of a residual module). Use of upsampling does not seem to have a significant improvement in performance.

Our upsampling scheme should not have any effect on subtasks A and C, where the worst class ratio is less than 2:1. Any difference is likely due to a different shuffling than their non-upsampling counterparts. The use of weighted loss seems to degrade the performance in tasks B and D. However, the difference is not significant.

6 Conclusion

We present FIMIF (Feature Interaction for Multimodal Integration and Fusion), a lightweight parameter-efficient model that leverages CLIP encoders for multimodal meme classification on PrideMM dataset. Our approach relies on aggressive dimensionality compression. A key finding from our ablation study is that the classification problem becomes mostly linear in nature after this compression, indicating that the dimensionality reduction itself is a critical component of our model's success. Our work highlights the potential of low-dimensional fusion as a viable path toward creating more efficient and sustainable models for complex multimodal tasks.

7 Limitations

Dependence On OCR Quality: The textual input relies heavily on the quality of the OCR. Errors in OCR, such as misread words or missing characters, are directly passed to the text encoder without correction or filtering. Moreover, CLIP's text encoder has a maximum context length of 77 tokens. This severely limits our model's ability to classify text-heavy memes. However, Table 5 indicates that the model achieves comparable performance even without OCR.

Lack Of Future Proofing: The world of memes on the Internet evolves rapidly. Words, images, and cultural references can shift in meaning over time. Since our model heavily relies on the frozen CLIP embeddings, it severely limits the ability of our model to adapt to emerging slangs, visual styles, and evolving socio-cultural contexts. This static representation may cause the model's performance to degrade over time.

8 Ethical Considerations

Environmental Impact: Training deep learning models can have a significant environmental impact, mainly due to high energy consumption and the resulting carbon emissions. To address this, we designed our model with a very low parameter count, which helps reduce the overall computational load. In practice, the most time-consuming step is the extraction of CLIP embeddings, while the actual training phase is relatively quick and lightweight. Our fine-tuning approach helps the model adapt quickly to new datasets, reducing the need for repeated or prolonged training.

Potential For Misuse: Any technology designed to understand and identify a specific type of content can potentially be used for malicious purposes. A model that learns the constituent elements of hateful memes could be used to generate new, more effective hateful content to systematically find loopholes in other detection systems.

Societal Impact Of Automated Moderation: The integration of automated moderation systems into digital platforms introduces several ethical concerns with severe societal implications. While such systems enable scalable and timely identification of harmful content, they also risk amplifying existing biases and disproportionately impacting certain user groups (Thapa et al., 2025b). Models trained on imbalanced or culturally narrow datasets may inadvertently silence marginalised communities, misclassify context-dependent expressions, or fail to generalise across linguistic and cultural boundaries. Automated moderation often lacks transparency and interpretability, limiting users' ability to understand or contest moderation decisions. This opacity can undermine fairness and accountability, particularly in high-stakes environments where content removal may affect public discourse or individual reputation.

References

Aashish Bhandari, Siddhant B Shah, Surendrabikram Thapa, Usman Naseem, and Mehwish Nasim. 2023. Crisishatemm: Multimodal analysis of directed and undirected hate speech in text-embedded images from russia-ukraine conflict. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1994–2003.

Pete Burnap and Matthew L. Williams. 2015. Cyber hate speech on twitter: An application of machine classification and statistical modeling for policy and decision making. *Policy & Internet*, 7(2):223–242.

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. pages 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, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. 2017. Automated hate speech detection and the problem of offensive language. *Proceedings of the International AAAI Conference on Web and Social Media*, 11.

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, Minneapolis, Minnesota. Association for Computational Linguistics.

Ona de Gibert, Naiara Perez, Aitor García-Pablos, and Montse Cuadros. 2018. Hate speech dataset from a white supremacy forum. In *Proceedings of the 2nd Workshop on Abusive Language Online (ALW2)*, pages 11–20, Brussels, Belgium. Association for Computational Linguistics.

Raul Gomez, Jaume Gibert, Lluis Gomez, and Dimosthenis Karatzas. 2020. Exploring hate speech detection in multimodal publications. In *2020 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1459–1467.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2015. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV)*, ICCV '15, page 1026–1034, USA. IEEE Computer Society.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2023. Debertav3: Improving deberta using electra-style pretraining with gradient-disentangled embedding sharing.

Niklas Heim, Tomas Pevny, and Vasek Smidl. 2020. Neural power units. In *Advances in Neural Information Processing Systems*, volume 33, pages 6573–6583. Curran Associates, Inc.

Ali Hürriyetoğlu, Surendrabikram Thapa, Hristo Tanev, and Surabhi Adhikari. 2025. Findings and insights from the 8th workshop on challenges and applications of automated extraction of socio-political events from text. In *Proceedings of the 8th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text (CASE 2025)*.

Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. 2021. Scaling up visual and vision-language representation learning with noisy text supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 4904–4916. PMLR.

Brendan Kennedy, Mohammad Atari, Aida M Davani, Leigh Yeh, Ali Omrani, Yehsong Kim, Kris Coombs, Gwenyth Portillo-Wightman, Shreya Havaldar, Elaine Gonzalez, and et al. 2022. The gab hate corpus.

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*, NIPS '20, Red Hook, NY, USA. Curran Associates Inc.

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.

Zhun Liu, Ying Shen, Varun Bharadhwaj Lakshminarasimhan, Paul Pu Liang, AmirAli Bagher Zadeh, and Louis-Philippe Morency. 2018. Efficient low-rank multimodal fusion with modality-specific factors. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2247–2256, Melbourne, Australia. Association for Computational Linguistics.

Andreas Madsen and Alexander Rosenberg Johansen. 2020. Neural arithmetic units. In *International Conference on Learning Representations*.

Binny Mathew, Punyajoy Saha, Seid Muhie Yimam, Chris Biemann, Pawan Goyal, and Animesh Mukherjee. 2021. Hatexplain: A benchmark dataset for explainable hate speech detection. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(17):14867–14875.

Anil Singh Parihar, Surendrabikram Thapa, and Sushruti Mishra. 2021. Hate speech detection using natural language processing: Applications and challenges. In *2021 5th International Conference on Trends in Electronics and Informatics (ICOEI)*, pages 1302–1308. IEEE.

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.

Jingnong Qu, Liunian Harold Li, Jieyu Zhao, Sunipa Dev, and Kai-Wei Chang. 2022. Disinfomeme: A multimodal dataset for detecting meme intentionally spreading out disinformation.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR.

Sara Rosenthal, Pepa Atanasova, Georgi Karadzhov, Marcos Zampieri, and Preslav Nakov. 2021. SOLID: A large-scale semi-supervised dataset for offensive language identification. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 915–928, Online. Association for Computational Linguistics.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2020. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter.

Daniel Schlör, Markus Ring, and Andreas Hotho. 2020. inalu: Improved neural arithmetic logic unit. *Frontiers in Artificial Intelligence*, Volume 3 - 2020.

Siddhant Bikram Shah, Shuvam Shiwakoti, Maheep Chaudhary, and Haohan Wang. 2024. Memeclip: Leveraging clip representations for multimodal meme classification. pages 17320–17332.

Rupesh Kumar Srivastava, Klaus Greff, and Jürgen Schmidhuber. 2015. Highway networks.

Surendrabikram Thapa, Farhan Ahmad Jafri, Ali Hürriyetoğlu, Francielle Vargas, Roy Ka Wei Lee, and Usman Naseem. 2023. Multimodal hate speech event detection-shared task 4. In *CASE 2023-Proceedings of the 6th Workshop on Challenges and Applications of Automated Extraction of Socio-Political Events from Text, associated with 14th International Conference on Recent Advances in Natural Language Processing, RANLP 2023*, pages 151–159. Association for Computational Linguistics.

Surendrabikram Thapa, Kritesh Rauniyar, Farhan Ahmad Jafri, Hariram Veeramani, Raghav Jain, Sandesh Jain, Francielle Vargas, Ali Hürriyetoğlu, and Usman Naseem. 2024. Extended multimodal hate speech event detection during russia-ukraine crisis-shared task at case 2024. In *7th Workshop on Challenges and Applications of Automated Extraction of Socio-Political Events from Text, CASE 2024*, pages 221–228. Association for Computational Linguistics.

Surendrabikram Thapa, Siddhant Bikram Shah, Kritesh Rauniyar, Shuvam Shiwakoti, Surabhi Adhikari, Hariram Veeramani, Kristina T. Johnson, Ali Hürriyetoğlu, Hristo Tanev, and Usman Naseem. 2025a. Multimodal hate, humor, and stance event detection in marginalized sociopolitical movements. In *Proceedings of the 8th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text (CASE 2025)*.

Surendrabikram Thapa, Shuvam Shiwakoti, Siddhant Bikram Shah, Surabhi Adhikari, Hariram Veeramani, Mehwish Nasim, and Usman Naseem. 2025b. Large language models (llm) in computational social science: prospects, current state, and challenges. *Social Network Analysis and Mining*, 15(1):1–30.

Robert Tibshirani. 1996. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288.

Andrew Trask, Felix Hill, Scott E Reed, Jack Rae, Chris Dyer, and Phil Blunsom. 2018. Neural arithmetic logic units. In *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc.

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.

Amir Zadeh, Minghai Chen, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. 2017. Tensor fusion network for multimodal sentiment analysis. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1103–1114, Copenhagen, Denmark. Association for Computational Linguistics.