TRACE: Textual Relevance Augmentation and Contextual Encoding for Multimodal Hate Detection

Girish A. Koushik[1*], **Helen Treharne**[2], **Aditya Joshi**[3], **Diptesh Kanojia**[1]
[1]Nature-Inspired Computing & Engineering, University of Surrey, Guildford, United Kingdom
[2]Surrey Centre for Cyber Security, University of Surrey, Guildford, United Kingdom
[3]University of New South Wales, Sydney, Australia
g.koushik@surrey.ac.uk, h.treharne@surrey.ac.uk, aditya.joshi@unsw.edu.au, d.kanojia@surrey.ac.uk

Abstract

Social media memes are a challenging domain for hate detection because they intertwine visual and textual cues into culturally nuanced messages. To tackle these challenges, we introduce **TRACE**, a hierarchical multimodal framework that leverages visually grounded context augmentation, along with a novel caption-scoring network to emphasize hate-relevant content, and parameter-efficient fine-tuning of CLIP's text encoder. Our experiments demonstrate that selectively fine-tuning deeper text encoder layers significantly enhances performance compared to simpler projection-layer fine-tuning methods. Specifically, our framework achieves state-of-the-art accuracy (0.807) and F1-score (0.806) on the widely-used Hateful Memes dataset, matching the performance of considerably larger models while maintaining efficiency. Moreover, it achieves superior generalization on the MultiOFF offensive meme dataset (F1-score 0.673), highlighting robustness across meme categories. Additional analyses confirm that robust visual grounding and nuanced text representations significantly reduce errors caused by benign confounders. We publicly release our code to facilitate future research.[1]

Code — https://github.com/surrey-nlp/TRACE
Extended version — https://arxiv.org/abs/2504.17902

1 Introduction

Social media platforms provide the environment by which multimedia content, such as internet memes, can proliferate, evolve swiftly, making them difficult to moderate (Young 2022). Internet memes represent benign, humorous, or satirical images combined with a caption in the form of overlaid text (Denisova 2019). In hateful memes, the images are repurposed to propagate hateful messages, reinforce stereotypes, and target individuals and communities. Detecting hateful memes requires not only an understanding of the visual content but also the underlying linguistic and cultural context that distinguishes hateful rhetoric from benign humour.

Figure 1: Illustration of benign confounders (absent from the dataset) as noted by Kiela et al. (2020). (left) meme, (centre) image confounder, (right) text confounder.

Understanding hateful memes poses two major challenges for current vision–language models. Firstly, memes are not simply a juxtaposition of an image with overlaid text; instead, they convey meaning through a nuanced fusion in which the image can provide context, irony, or subtext to the accompanying words in the text. Hence, models need to capture subtle semantic and cultural cues within mathematical representations. This remains difficult for the current state-of-the-art architectures (Kiela et al. 2020). Secondly, models often misclassify content because of the impact that benign elements have on representations. For example, in Figure 1 (top row, left-most image), the text along with the image yields a hateful meme. However, the benign image along with the same text (top row, middle image) yields a non-hateful meme. Such combinations are known as "benign confounders" (Kiela et al. 2020). While this is an example of a textual confounder, similar image confounders also exist. These confounders prevent models from exploiting a combination of unimodal representations, *i.e.*, neither the visual modality nor the textual modality alone reliably indicates the presence of hate (Aggarwal et al. 2024; Koushik, Kanojia, and Treharne 2025). As a result, there is a need for advanced multimodal strategies for the detection of hateful memes.

Existing approaches to hateful meme detection span a diverse range of multimodal techniques, each attempting to reconcile the conflicting and often subtle signals present in meme images and their overlaid text (Hu et al. 2024; Mei et al. 2023; Kumar and Nandakumar 2022). Many of these

*Corresponding author

[1]**Disclaimer**: This paper includes references to potentially disturbing, hateful, or offensive content due to the nature of the task.

approaches achieve good results, but they focus on either capturing the complex semantic interplay inherent in memes or mechanisms for distinguishing benign confounders. To understand memes, we believe a framework needs to encapsulate a holistic understanding of the image and text combination along with cultural context, which may be missing from the original meme text. To this end, we propose a hierarchical interpretable framework, **TRACE**, that performs *multimodal context augmentation*, while enriching input information related to the meme image prior to augmentation. The **main contributions** of this paper are:

- A hierarchical, interpretable framework, referred to as **TRACE**, for hate detection (as shown in Figure 2) that includes visual grounding to reduce the complexity of meme images when performing multimodal context augmentation.
- Within **TRACE**, we propose an efficient n-layer fine-tuning approach, and novel joint optimization of vision-language models using a caption-scoring neural network. The caption scorer takes candidate captions as input and selects the best image-caption pair during the fine-tuning process.
- A comprehensive evaluation of various model configurations with both quantitative and qualitative error analysis. Notably, **TRACE** attains high accuracy (0.807) and an F1 score (0.806) on the Hateful Memes dataset, at par with the existing SoTA framework while being more efficient.

2 Related Work

Multimodal Hate Detection Recent advances in multimodal hate detection have shown promising developments in analysing hateful memes, though significant challenges persist. Gomez et al. (2020) established early foundations with the MMHS150K dataset, implementing a dual-stream architecture combining Inceptionv3 (Szegedy et al. 2016) for visual features and LSTM (Hochreiter and Schmidhuber 1997) for textual features achieving a modest accuracy of 68.2%. The field then shifted toward contrastive learning approaches: HateSieve (Su et al. 2024) combined a large vision–language model (LVLM) for caption generation with SDXL (Meng et al. 2021) for image synthesis, achieving 73.45% accuracy on the FHM dataset (Kiela et al. 2020); Hate-CLIPper (Kumar and Nandakumar 2022) fine-tuned Contrastive Language-Image Pre-training (CLIP)'s projection layers for cross-modal interaction to reach an AUROC score of 0.858 (Acc 0.756), which was further improved by Retrieval-guided Contrastive Learning (RGCL) (Mei et al. 2023) to AUROC 0.870 (Acc 0.788) through a runtime retrieval database. While these methods outperform earlier models, they still struggle with benign confounders and complex semantic relationships between modalities, as cosine-based similarity can yield inconsistent performance (Steck, Ekanadham, and Kallus 2024). More compute-intensive approaches like PALI-X-VPD (Hu et al. 2024) achieve state-of-the-art AUROC 0.892 (Acc 0.808) via a $55B$-parameter language model with chain-of-thought prompting, raising concerns about real-world deployability. Alternative lightweight architectures such as fine-tuned Flamingo (Alayrac et al. 2022) and ISSUES (Burbi et al. 2023) have reported competitive AUROC scores of 0.866 and 0.855, respectively. Domain-specific datasets have also emerged: MultiOFF (Suryawanshi et al. 2020) for the 2016 US elections (F1 0.54) and CrisisHateMM (Bhandari et al. 2023) for Russia–Ukraine conflict memes (F1 0.786).

Building on these foundations, recent work has explored richer semantic and knowledge-guided representations. (Zhong and Baghel 2024) introduces a fairness-aware vision–language framework that generates human-centric explanations for multimodal memes across domains, focusing on interpretability rather than classification accuracy. (Grasso et al. 2024) presents KERMIT, a memory-augmented model integrating ConceptNet knowledge for harmful meme detection, reporting 85.3% AUROC on the FHM dataset. (Lin et al. 2024) harnesses multimodal debates between large language models to generate conflicting rationales, then fine-tunes a lightweight judge model for harmfulness inference, improving explainability and detection robustness. These approaches suggest that integrating external knowledge, richer semantic signals, and fairness considerations can complement contrastive and large-scale models, balancing detection accuracy with interpretability and deployment feasibility.

Caption Generation The development of vision-language models (VLMs) has revolutionized image captioning and expansion, producing more detailed, context-rich descriptions. Early encoder-decoder models like "Show and Tell" (Vinyals et al. 2015) and "Show, Attend and Tell" (Xu et al. 2015) used convolutional and recurrent neural networks, but typically generated only brief captions. Recent models use transformers and large pretrained language models for improved semantic alignment; ClipCap (Mokady, Hertz, and Bermano 2021), for example, augments CLIP embeddings with a language model for captions closely tied to image content. Large Vision-Language Models (LVLMs) like BLIP-2 (Li et al. 2023) and InternVL (Chen et al. 2024b) have achieved state-of-the-art results on captioning tasks. In our work, we use the INTERNVL-2.5-8B model for caption generation, as the meme text forms meaning in tandem with the image rather than describing it directly.

Visual Grounding Despite LVLMs' strong image understanding, they can hallucinate on subtle meme content. Supplementing them with visual grounding provides needed context and improves model interpretation (Bai et al. 2024). Recent advances integrate visual grounding into VLMs, boosting alignment between text and visual regions, benefitting tasks like VQA, image retrieval, and object localization. Notably, some works rephrase and augment input queries by salient visual content (Prasad, Stengel-Eskin, and Bansal 2023), while others use generative VLM prompts to elicit object-level descriptions without extensive manual labeling (Wang et al. 2025). Additional methods apply visual grounding for object localization (Yang et al. 2023), image compression (Liu et al. 2024a), and reducing hallucinations (Yan et al. 2024). Accordingly, our approach leverages visual grounding through object tagging and detection to im-

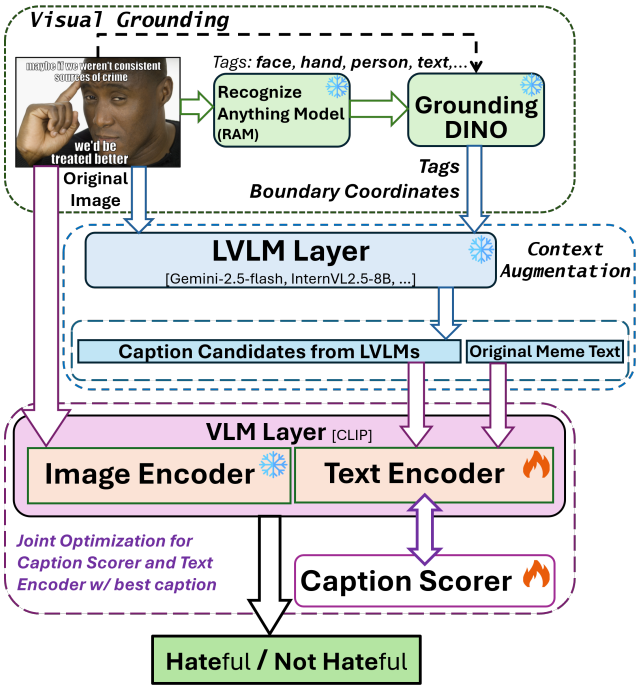

Figure 2: Proposed hierarchical architecture of the **TRACE** framework for Hateful Meme detection. The process includes: (1) Visual Grounding using RAM and GroundingDINO; (2) Context Augmentation where LVLMs generate captions incorporating grounded visual details and original text; (3) A VLM layer (such as CLIP) encoding the image and augmented text; (4) A novel Caption Scorer selecting the most relevant caption; and (5) Joint fine-tuning the text encoder and caption scorer for final classification. Frozen components (❄) remain untrained.

prove caption quality and minimize hallucination.

3 Methodology

Our **TRACE** framework consists of three blocks as shown in Figure 2. We describe the first two blocks in § 3.1, and the third block in § 3.2. Figure 3 provides a more detailed view of the third block. In both figures, the ❄ symbol indicates that framework components in **TRACE** are frozen and thus remain untrained on our data. The 🔥 symbol highlights that **TRACE** proposes training for these components.

3.1 Visually Grounded Context Augmentation

To create visual grounding from the original image, we utilise the Recognize Anything Model (RAM) (Zhang et al. 2024) for tagging, and GroundingDINO (Liu et al. 2024b) for detecting objects along with their bounding boxes. Although zero-shot object detection methods, *e.g.*, YOLOv9 (Wang, Yeh, and Mark Liao 2024), can produce bounding boxes for generic images, they fall short in capturing the context of hateful memes. Hence, we employ RAM for tag generation paired with GroundingDINO for open-vocabulary object detection. The next step is to augment the

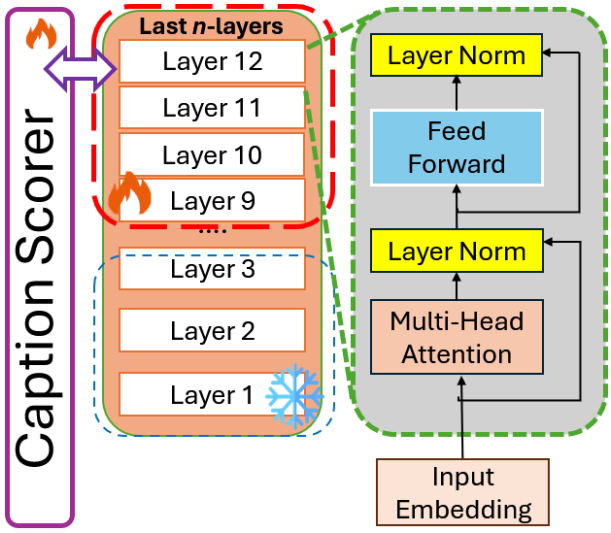

Figure 3: Parameter-Efficient Fine-tuning of the CLIP text encoder within **TRACE**. This diagram illustrates the strategy where only the last n-layers (indicated by 🔥 symbol) of the text encoder (comprising Layer Norm, Multi-Head Attention, and Feed Forward blocks) are fine-tuned, alongside the Caption Scorer. This minimizes trainable parameters to prevent overfitting while adapting the model for hate detection.

context from the image into the text. To do so, the tags from RAM and bounding box coordinates from GroundingDINO are then fed into the LVLM layer to generate captions. These captions highlight interactions between visual and textual elements, with an emphasis on cultural references essential for meme comprehension, as shown in the supplementary file[2]. **TRACE** supports the generation of captions using several different VLMs. All such captions and the original meme text are then encoded for the caption scorer, as described below.

3.2 Caption-aware Text Encoder Fine-tuning

The last block of **TRACE** is a combination of a vision-language model and a novel neural network. **TRACE** builds upon CLIP, a foundational vision-language encoder model which was trained using a contrastive learning approach (Radford et al. 2021). CLIP is jointly trained using image-text pairs, and our task requires an understanding of the nuanced interplay between visual content and text for detecting hateful memes, making it well-suited as a component. Since our training dataset contains limited samples (approx. $8.5k$), fine-tuning all layers in the text encoder (as shown in Table 1) is likely to result in over-fitting on the training set, worsening performance on the validation and the test set. Therefore, for such a low-resource scenario, **TRACE** adopts parameter-efficient fine-tuning (Houlsby et al. 2019), and proposes only tuning the last n-layers of the text encoder, as shown in Figure 3. Additionally, our experiments show the viability of only tuning

[2]All references to the supplementary file points to the extended version of our paper on arXiv

Model	# of Params	# of Text Enc. Layers	Text Enc.	# Last-4
CLIP-ViT-L/14	428M	12	123M	41M
SIGLIP2-L/16-384	882M	24	413M	68M
CLIP-XLM-R-ViT-H-14	1.1B	24	540M	90M

Table 1: Number (#) of parameters for CLIP and SigLIP models. 'Text Enc.' reports parameters from the text encoder, and 'Last-4' reports parameters in the last four layers of the text encoder.

the text encoder, leaving the image encoder untrained.

TRACE feeds the candidate captions to the text encoder for obtaining representations, and further proposes the use of our novel caption scorer, *i.e.*, a feed-forward neural network with 3 hidden layers detailed as follows. The **caption scorer** $\mathcal{S}$ processes textual features $h_i \in \mathbb{R}^d$ (where d is the output dimension of the CLIP text encoder) to produce relevance scores S_i.

$$S_i = f_\theta(\mathbf{h}_i) = \mathbf{W}_5\Big(\phi_4\Big(\mathbf{W}_4\Big(\phi_3\Big(\mathbf{W}_3\Big(\phi_2\big(\mathbf{W}_2\big(\phi_1(\mathbf{W}_1\mathbf{h}_i)\big)\big)\Big)\Big)\Big)\Big)\Big)$$

where:

$$\phi(x) = \text{GELU}(\text{LayerNorm}(x)) \odot \text{Dropout}(p)$$

$$\mathbf{W}_1 \in \mathbb{R}^{d\times 1024}, \quad \mathbf{W}_2, \mathbf{W}_3 \in \mathbb{R}^{1024\times 1024}, \mathbf{W}_4 \in \mathbb{R}^{1024\times 512}, \quad \mathbf{W}_5 \in \mathbb{R}^{512\times 1}$$

Weight normalisation $\mathbf{W} = \frac{g}{\|\mathbf{v}\|}\mathbf{v}$ is applied to the weight matrices $\mathbf{W}_2$, $\mathbf{W}_3$, and $\mathbf{W}_4$. Bias terms are omitted for brevity, but applied in each linear layer. Second, **Gumbel-Softmax** enables differentiable caption selection:

$$p_i = \frac{exp((s_i + g_i)/\tau}{\Sigma_{j=1}^{n} exp((s_j + g_j)/\tau}$$

where $g_i \sim -log(-log(\mathcal{U}(0,1)))$ is Gumbel noise, τ is the temperature, and n is the number of caption candidates. The selected caption and image features are projected onto a larger projection space and then undergo bidirectional cross-attention fusion before classification. Given image features $\mathbf{I}$ and text features $\mathbf{T}$, the cross-attention mechanism computes:

$$\mathbf{I}_p = \text{Projection}(\mathbf{I}) \in \mathbb{R}^{B\times D}$$
$$\mathbf{T}_p = \text{Projection}(\mathbf{T}) \in \mathbb{R}^{B\times D}$$

where B is batch size and D is projection dimension (1024).

$$\text{CrossAttn}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{d_K}}\right)\mathbf{V}$$

For image-to-text attention:

$$\mathbf{I}_{\text{enhanced}} = \mathbf{I_p} + \text{CrossAttn}(\mathbf{I_p}, \mathbf{T_p}, \mathbf{T_p})$$

where $\mathbf{Q} = \mathbf{I}_p$ (Query is image), $\mathbf{K} = \mathbf{T}_p$ (Key is text), and $\mathbf{V} = \mathbf{T}_p$ (Value is text). For text-to-image attention:

$$\mathbf{T}_{\text{enhanced}} = \mathbf{T_p} + \text{CrossAttn}(\mathbf{T_p}, \mathbf{I_p}, \mathbf{I_p})$$

where $\mathbf{Q} = \mathbf{T}_p$ (Query is text), $\mathbf{K} = \mathbf{I}_p$ (Key is image), and $\mathbf{V} = \mathbf{I}_p$ (Value is image). The enhanced features are concatenated:

$$\mathbf{F}_{\text{combined}} = [\mathbf{I}_{\text{enhanced}}; \mathbf{T}_{\text{enhanced}}] \in \mathbb{R}^{B\times 2D}$$

This bidirectional attention allows each modality to be enhanced by information from the other before classification. Finally, a **hate relevance loss** $\mathcal{L}_{\text{rel}}$ aligns caption selection with labels:

$$\mathcal{L}_{\text{rel}} = -\left[y \log\left(\sum p_i\right) + (1-y)\log\left(1 - \sum p_i\right)\right]$$

The classification loss ($\mathcal{L}_{\text{cls}}$) uses combined features of both the image and the selected caption. Hate relevance loss ($\mathcal{L}_{\text{rel}}$) on the other hand, uses only the caption scores directly from the caption scorer. Both use *binary cross-entropy loss* for calculation. While $\mathcal{L}_{\text{cls}}$ predicts whether an image-text pair represents hateful content, $\mathcal{L}_{\text{rel}}$ directly evaluates how 'hateful' each caption is and encourages the caption scorer to assign higher scores to hateful captions for hateful images. These components interact through joint optimization:

$$\mathcal{L}_{\text{total}} = \underbrace{\mathcal{L}_{\text{cls}}}_{\text{classification}} + \underbrace{\mathcal{L}_{\text{rel}}}_{\text{hate alignment}}$$

The model learns to: 1) weight captions by hate relevance ($\mathcal{S}$), 2) maintain differentiability in selection (p_i), and 3) directly connect caption choices to label supervision ($\mathcal{L}_{\text{rel}}$), enabling the model to specialize for hate detection while preserving original text understanding through partial parameter updates. These losses are back-propagated to the model for fine-tuning.

3.3 Experimental Setup

We perform experiments in three different settings as described below. For generalizability, we also report the performance of the best approach on the MultiOFF dataset. Both datasets are binary classification datasets with split sizes as shown in Table 2.

Zero-shot We perform zero-shot experiments with multiple LVLMs for hate detection, allowing us to report the performance of only the LVLM layer, without any other component of **TRACE**. For these experiments, we provide the original meme image and overlay text as input within the prompt described in the supplementary file.

Dataset	#Training	#Val	#Test
Hateful Memes	$8.5k$	500	$1k$
MultiOFF	445	149	149

Table 2: Dataset Splits. Number of samples allocated to the training, validation (Val), and test sets for the Facebook Hateful Memes (FHM) and MultiOFF datasets used in the experiments.

Projection Layer Fine-tuning This experiment compares the performance of two VLMs, CLIP-ViT-L/14, and LLM2CLIP-Llama-3.2 (1B-Instruct-CC-Finetuned) with two existing approaches, HateCLIPper-Align (Kumar and Nandakumar 2022), and RGCL (Mei et al. 2023) which utilize representation fusion. For fine-tuning both VLMs, we first take the encoder pooler outputs (*i.e.*, the [CLS] token) and map them through linear plus dropout blocks for text and image separately. These mapped representations are then passed through a short stack of "pre-output" layers, which apply dropout, linear transformation, and an activation function before the final classification.

TRACE We utilize InternVL-2.5-8B, and Gemini-2.5-flash to generate candidate captions for **TRACE**. We also perform multiple experiments varying VLM models, and the number of tunable layers (n) in the final framework block, while jointly optimizing the caption scorer for hate detection. We perform experiments with CLIP-ViT-L/14, SigLIP2-L/16-384 (Tschannen et al. 2025) and the large variant CLIP-XLM-R-ViT-H/14 (Cherti et al. 2023) models as they report strong performance on downstream tasks involving vision-language.

All publicly available models are obtained from the HuggingFace repository, while for Gemini-2.5-flash and GPT-4o-mini, we use their respective APIs. The experiments are conducted using two $24GB$ $A5000$ GPUs. All experiments on both datasets (Table 2) are executed with a batch size of 64, with a gradient accumulation up to 512, and a learning rate of $1e-4$ for 30 epochs with early stopping, with an average runtime of 2 hours per experiment.

Evaluation Metrics To evaluate our models' ability to flag hateful memes, we report macro-averaged precision, recall, F1-score, and accuracy using the Scikit-learn toolkit (Kramer and Kramer 2016). We omit AUROC score, since it only measures the ranking of predicted scores and ignores whether those probabilities reflect true hateful content (Lobo, Jiménez-Valverde, and Real 2008). Furthermore, it aggregates performance across all possible thresholds, rather than relying on a single decision point used in deployment. It treats false positives and false negatives as equally important, despite missed hateful instances (false negatives) being far more problematic than over-flagging benign memes (Halligan, Altman, and Mallett 2015). Consequently, F1 and accuracy offer more meaningful insights into our classifiers' real-world effectiveness.

4 Experimental Results

Zero-shot From Table 3, the zero-shot results show that Gemini-2.5-flash (Team et al. 2023) performs significantly better than the other two LVLMs, achieving an F1 score of 0.756, while the 8B variant of the InternVL-2.5-8B model (Chen et al. 2024a) achieves 0.679, and GPT-4o-mini (Hurst et al. 2024) achieves 0.703 F1. To evaluate the viability of LLMs for such a task, we further experimented by providing the generated captions from LVLMs as input to an LLM for zero-shot hate detection. However, the results were significantly inferior to LVLMs in zero-shot.

Model	Acc.	F1	P	R
InternVL-2.5-8B	0.690	0.679	0.688	0.671
Gemini-2.5-flash	0.741	0.756	0.698	0.824
GPT-4o-mini	0.692	0.703	0.666	0.745

Table 3: Zero-Shot hate detection performance on the FHM Dataset. Evaluates the performance of various LVLMs on the FHM test split without any task-specific fine-tuning. Acc: Accuracy, P: Precision, R: Recall.

Model	Acc.	F1	P	R
CLIP-ViT-L/14	0.720	0.710	0.746	0.717
LLM2CLIP-Llama-3.2	0.673	0.652	0.716	0.668
Hate-CLIPper - Align	$0.756^{\star}$	–	–	–
RGCL	0.788	–	–	–

Table 4: Projection layer fine-tuning results on the FHM Dataset. Compares performance when fine-tuning only the projection layers of CLIP-ViT-L/14 and LLM2CLIP-Llama-3.2-1B. Results are benchmarked against prior methods (Hate-CLIPper-Align, RGCL) using similar lightweight tuning. *Score reproduced from their code.

Projection Layer Fine-tuning As shown in Table 4, although the CLIP-ViT-L/14 and the newer variant of LLM2CLIP (Huang et al. 2024) models already report good results in other vision–language tasks, simply fine-tuning the projection layers on the generated captions did not suffice. They achieve accuracy scores of 0.720 and 0.673, respectively; however, they still fall short compared to existing CLIP-based approaches such as Hate-CLIPper (Acc 0.788) and RGCL (Acc 0.756). This suggests that simple projection layer fine-tuning is insufficient for effectively utilizing the caption information.

TRACE's Fine-tuning As Table 5 demonstrates, selectively fine-tuning the text encoder yields a clear performance boost compared to full text-encoder fine-tuning in smaller models. In particular, increasing the number of trainable layers to 4 yields the best F1 score, as visualized in Figure 4. This pattern appears for CLIP-ViT-L/14, SigLIP2-L/16-384, and CLIP-XLM-R-ViT-H/14, indicating that deeper text encoder updates improve the model's ability to capture subtle, context-dependent signals in hateful memes for CLIP-based models. With the proposed hate-relevance loss, we observe **TRACE** achieves SoTA performance on the FHM dataset, with an F1 of 0.806 and an Accuracy of 0.807. Please refer to Appendix D in the supplementary file for the statistical significance (Pairwise McNemar's) tests.

Finally, as shown in Table 6, we also evaluated **TRACE** (w/ CLIP-XLM-R-ViT-H-14) with InternVL-2.5-8B and Gemini-2.5-flash on the MultiOFF dataset to demonstrate the generalizability of our approach. The table also presents the best existing benchmarked results as presented in the original dataset paper for MultiOFF (Baseline). Despite the smaller sample size and different domain focus

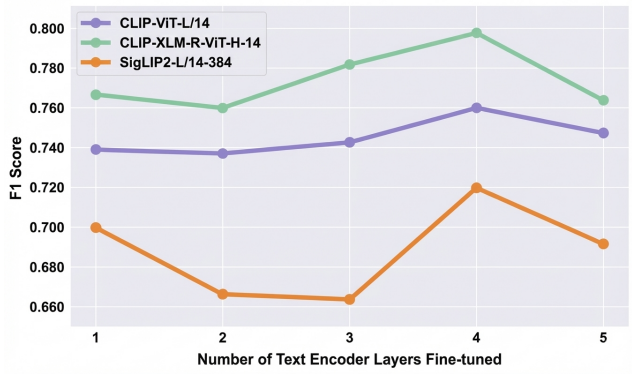

Figure 4: F1 Score vs. the number of fine-tuned text encoder layers (n). Illustrates the impact of increasing the number of trainable layers (n, from 1 to 5) in the text encoder on F1 performance for CLIP-ViT, CLIP-XLM, and SigLIP models within the **TRACE** framework on the FHM dataset. Fine-tuning 4 layers achieves the best performance for all three models.

of MultiOFF, our framework still achieves a notably higher F1 score compared to prior work. This result underlines the broader applicability of our methodology, even when dealing with more limited datasets or varied offensive content scenarios.

4.1 Discussion

Overall, **TRACE** performs at par with the PALI-X-VPD (Hu et al. 2024) framework, which currently reports SoTA performance over the FHM dataset. It reports an accuracy of 0.808, while **TRACE** reports an accuracy of 0.807 and an F1 score of 0.806. Further analysis using last-layer fine-tuning (Table 5) isolates the impact of the loss components under minimal parameter updates. This suggests that our novel caption scorer and the associated $\mathcal{L}_{\text{rel}}$ provide targeted and effective signals to adapt the model to this specific hate detection task.

Importance of Caption-Scorer and Visual Grounding Our qualitative analysis highlights the critical roles of visual grounding and caption scoring in **TRACE**. Visual grounding notably reduces hallucinations in smaller models (e.g., INTERNVL-2.5-8B), aiding accurate entity identification. For instance, in meme '07193.png' (Figure 5a), visual grounding improved INTERNVL-2.5-8B's basic visual recognition but still missed subtle racial stereotypes implied by the text "things I love to hunt." In contrast, the larger GEMINI-2.5-FLASH model generated a caption explicitly recognizing these racial stereotypes, which the caption scorer correctly identified as most hate-relevant (shown in Appendix B of the supplementary file). This targeted caption selection enabled **TRACE** to effectively detect nuanced hate, demonstrating how the combination of robust visual grounding and the caption scorer significantly enhances multimodal hate detection accuracy and interpretability.

Model	FT	Acc.	F1	P	R
CLIP-VIT L/14	T (-1)	0.743	0.740	0.757	0.745
	T (-4)	0.764	0.764	0.764	0.764
SIGLIP2 L/16-384	T (-1)	0.705	0.700	0.730	0.708
	T (-4)	0.723	0.721	0.733	0.725
CLIP-XLM-T VIT-H/14	T (-1)	0.774	0.771	0.794	0.776
	T (-4)	**0.807**	**0.806**	**0.813**	**0.808**

Table 5: Performance of **TRACE** fine-tuning on the FHM Dataset with ablations. Evaluates **TRACE** using different VLMs and varying the number of fine-tuned text encoder layers ($n = 1$ and $n = 4$) using $\mathcal{L}_{\text{cls}} + \mathcal{L}_{\text{rel}}$ losses. The best configuration achieving SoTA-comparable results is marked in bold. FT: Fine-tuning, T: Text, Acc: Accuracy, P: Precision, R: Recall. (-1) and (-4) indicate that last layer and last 4 layers of the text encoder are fine-tuned, respectively.

Model Configuration	Acc.	F1	P	R
MultiOFF Baseline (Suryawanshi et al. 2020)	–	0.540	0.390	**0.840**
GEMINI-2.5-FLASH (Zero-shot)	0.557	0.492	0.451	0.542
CLIP-VIT-L/14 (projection FT)	0.604	0.377	0.302	0.500
TRACE (ours)	**0.678**	**0.673**	**0.673**	0.681

Table 6: Performance comparison on the MultiOFF Dataset. Evaluates the generalizability of **TRACE** by comparing its performance against the original MultiOFF baseline, zero-shot, and projection layer fine-tuning (FT). **TRACE** results use the configuration CLIP-XLM-R-VIT-H/14, and $n = 4$ layers.

Qualitative Analysis We further evaluated the **TRACE**'s best output from the CLIP-XLM-R-VIT-H/14 model by conducting qualitative analysis on samples with benign confounders. Figure 5 shows a confounder pair where the same template can be benign or hateful depending on context. **TRACE** correctly classifies both memes. Visual grounding focuses captioning on the salient entities (*e.g.,* the runner and overlaid text), and the caption scorer prioritizes the more nuanced caption. Figure 6 contrasts a correct negative with a false negative. In 6a, a common, non-hateful template is correctly rejected. In 6b, the meme implies potential violence, but small, low-salience objects (guns) are not consistently captured by the grounding stage, leading the classifier to mislabel. This points to a remaining limitation: sensitivity to small or partially occluded objects that carry decisive semantic weight. These examples show that **TRACE** is robust to benign confounders when salient cues are grounded, but can miss cases where key evidence is small or visually subtle. The supplementary file presents additional examples predicted by **TRACE**, comparing different candidate captions, their scores, and how these eventually influence the classification outcome.

(a) True Label: Hateful, Prediction: Hateful

(b) True Label: Not Hateful, Prediction: Not Hateful

Figure 5: **TRACE** predictions using CLIP-XLM. (a) A correct classification where subtle hateful objectification is identified. (b) A correctly identified non-hateful meme.

(a) True Label: Not Hateful, Prediction: Not Hateful

(b) True Label: Hateful, Prediction: Not Hateful

Figure 6: **TRACE** predictions using CLIP-XLM. (a) A correctly identified non-hateful meme. (b) A misclassification where a hateful meme implying potential violence (guns) is missed, possibly due to visual grounding challenges with small objects.

TRACE as an Interpretable Framework Existing approaches, such as HateCLIPper (Kumar and Nandakumar 2022) and RGCL (Mei et al. 2023), only produce classification labels, and there is no way for a moderator to understand how the model arrived at that decision. **TRACE** 's modular design makes use of visual grounding, caption generation, and Gumbel-softmax reliant caption scorer to assign probabilities to any candidate captions. This helps the CLIP model pick the best assisting caption to make an informed prediction. Thus, when a moderator looks at the predicted label along with the caption scores and probabilities, they can understand the model's decision process regardless of a correct or incorrect model prediction.

Computational Efficiency To further highlight **TRACE** 's computational efficiency and methodological novelty, we compare it directly against PaLI-X-VPD (Hu et al. 2024), the current SoTA approach. As shown in Table 7, **TRACE** achieves comparable accuracy to PaLI-X-VPD while requiring drastically fewer computational resources, significantly reducing training time from two days using large TPU clusters to approximately four hours on just two GPUs while incorporating visual grounding models and LVLMs for generating contextually enriched captions. These components not only enhance detection accuracy but also provide interpretability by clearly linking multimodal signals to model predictions. Such efficiency and interpretability advantages position **TRACE** favorably for practical deployment scenarios, especially in resource-constrained environments.

Metric	PaLI-X-VPD (Hu et al. 2024)	TRACE (ours)
Model Size	$55B$ parameters + code gen + visual tools + CoT	$1.1B$ parameters (largest variant) + vg (RAM + GroundingDINO) + cap gen (INTERNVL-2.5-8B + GEMINI-2.5-FLASH)
Hardware Used	128 TPU-v3 / 128 TPU-v5	2 × NVIDIA A5000 GPUs (24GB each)
Training Time	∼2 days	∼4 hours (2h cap gen + 2h training)
Inference Time (avg)	8.9s (4.7s code gen. + 4.2s exec.)	∼8s (5.5s vg & cap gen + 2.5s single forward pass)
Acc (FHM)	0.808	0.807
F1 (FHM)	-	0.806

Table 7: Efficiency comparison with PaLI-X-VPD

5 Conclusion and Future Work

In conclusion, we propose the **TRACE** framework that enriches hateful meme detection with visually grounded context augmentation, caption scoring, and parameter-efficient text encoder fine-tuning. Our experiments showed that simply fine-tuning projection layers on strong underlying models does not fully leverage the generated captions. Instead, targeted text encoder tuning substantially improved accuracy and F1 across datasets (FHM and MultiOFF). Notably, our performance is at par with the $55B$ parameter SoTA PALI-X-VPD framework and **TRACE** achieves an accuracy of 0.807 and an F1 score of 0.806 while being more efficient. This reflects the framework's ability to capture subtle linguistic signals, an ability reinforced by visual grounding and enhancing context with LVLMs.

Future Work Although our text encoder fine-tuning strategy already shows strong performance, it still involves tuning millions of parameters, suggesting that scaling up the training set—*e.g.*, incorporating larger datasets such as MMHS150K or combining multiple (recently released) meme datasets could substantially enhance both generalizability and robustness. In addition, we plan to explore the extraction of intermediate-layer representations from encoders, since different layers may capture distinct linguistic or semantic nuances. Notably, all these directions fit naturally into our proposed framework: by treating it as the core pipeline, such refinements remain modular to apply to any meme dataset.

Ethical Impact

Our work relies on public datasets and publicly available pre-trained models; therefore, no ethical review was necessary. Nevertheless, these models can inherit biases from their original training data, potentially yielding skewed or harmful judgments for under-represented groups. From an ethical standpoint, responsible usage of this system would require human oversight to prevent over-censorship of benign content and to ensure a review of sensitive or ambiguous cases.

References

Aggarwal, P.; Mehrabanian, J.; Huang, W.; Alaçam, Ö.; and Zesch, T. 2024. Text or Image? What is More Important in Cross-Domain Generalization Capabilities of Hate Meme Detection Models? In *Findings of the Association for Computational Linguistics: EACL 2024*, 104–117.

Alayrac, J.-B.; Donahue, J.; Luc, P.; Miech, A.; Barr, I.; Hasson, Y.; Lenc, K.; Mensch, A.; Millican, K.; Reynolds, M.; et al. 2022. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35: 23716–23736.

Bai, Z.; Wang, P.; Xiao, T.; He, T.; Han, Z.; Zhang, Z.; and Shou, M. Z. 2024. Hallucination of multimodal large language models: A survey. *arXiv preprint arXiv:2404.18930*.

Bhandari, A.; Shah, S. B.; Thapa, S.; Naseem, U.; and Nasim, M. 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*, 1993–2002.

Burbi, G.; Baldrati, A.; Agnolucci, L.; Bertini, M.; and Del Bimbo, A. 2023. Mapping Memes to Words for Multimodal Hateful Meme Classification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2832–2836.

Chen, Z.; Wang, W.; Cao, Y.; Liu, Y.; Gao, Z.; Cui, E.; Zhu, J.; Ye, S.; Tian, H.; Liu, Z.; et al. 2024a. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *arXiv preprint arXiv:2412.05271*.

Chen, Z.; Wu, J.; Wang, W.; Su, W.; Chen, G.; Xing, S.; Zhong, M.; Zhang, Q.; Zhu, X.; Lu, L.; et al. 2024b. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 24185–24198.

Cherti, M.; Beaumont, R.; Wightman, R.; Wortsman, M.; Ilharco, G.; Gordon, C.; Schuhmann, C.; Schmidt, L.; and Jitsev, J. 2023. Reproducible scaling laws for contrastive language-image learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2818–2829.

Denisova, A. 2019. *Internet memes and society: Social, cultural, and political contexts*. Routledge.

Gomez, R.; Gibert, J.; Gomez, L.; and Karatzas, D. 2020. Exploring hate speech detection in multimodal publications. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, 1470–1478.

Grasso, B.; La Gatta, V.; Moscato, V.; and Sperlì, G. 2024. Kermit: Knowledge-empowered model in harmful meme detection. *Information Fusion*, 106: 102269.

Halligan, S.; Altman, D. G.; and Mallett, S. 2015. Disadvantages of using the area under the receiver operating characteristic curve to assess imaging tests: a discussion and proposal for an alternative approach. *European radiology*, 25(4): 932–939.

Hochreiter, S.; and Schmidhuber, J. 1997. Long short-term memory. *Neural computation*, 9(8): 1735–1780.

Houlsby, N.; Giurgiu, A.; Jastrzebski, S.; Morrone, B.; De Laroussilhe, Q.; Gesmundo, A.; Attariyan, M.; and Gelly, S. 2019. Parameter-efficient transfer learning for NLP. In *International conference on machine learning*, 2790–2799. PMLR.

Hu, Y.; Stretcu, O.; Lu, C.-T.; Viswanathan, K.; Hata, K.; Luo, E.; Krishna, R.; and Fuxman, A. 2024. Visual program distillation: Distilling tools and programmatic reasoning into vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9590–9601.

Huang, W.; Wu, A.; Yang, Y.; Luo, X.; Yang, Y.; Hu, L.; Dai, Q.; Dai, X.; Chen, D.; Luo, C.; et al. 2024. Llm2clip: Powerful language model unlock richer visual representation. *arXiv preprint arXiv:2411.04997*.

Hurst, A.; Lerer, A.; Goucher, A. P.; Perelman, A.; Ramesh, A.; Clark, A.; Ostrow, A.; Welihinda, A.; Hayes, A.; Radford, A.; et al. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

Kiela, D.; Firooz, H.; Mohan, A.; Goswami, V.; Singh, A.; Ringshia, P.; and Testuggine, D. 2020. The hateful memes challenge: Detecting hate speech in multimodal memes. *Advances in neural information processing systems*, 33: 2611–2624.

Koushik, G. A.; Kanojia, D.; and Treharne, H. 2025. Towards a Robust Framework for Multimodal Hate Detection: A Study on Video vs. Image-based Content. In *Companion Proceedings of the ACM on Web Conference 2025*, 2014–2023.

Kramer, O.; and Kramer, O. 2016. Scikit-learn. *Machine learning for evolution strategies*, 45–53.

Kumar, G. K.; and Nandakumar, K. 2022. Hate-CLIPper: Multimodal hateful meme classification based on cross-modal interaction of CLIP features. *arXiv preprint arXiv:2210.05916*.

Li, J.; Li, D.; Savarese, S.; and Hoi, S. 2023. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, 19730–19742. PMLR.

Lin, H.; Luo, Z.; Gao, W.; Ma, J.; Wang, B.; and Yang, R. 2024. Towards explainable harmful meme detection through multimodal debate between large language models. In *Proceedings of the ACM Web Conference 2024*, 2359–2370.

Liu, J.; Wei, Y.; Lin, J.; Zhao, S.; Sun, H.; Chen, Z.; Zeng, W.; and Jin, X. 2024a. Tell Codec What Worth Compressing: Semantically Disentangled Image Coding for Machine with LMMs. In *2024 IEEE International Conference on Visual Communications and Image Processing (VCIP)*, 1–5. IEEE.

Liu, S.; Zeng, Z.; Ren, T.; Li, F.; Zhang, H.; Yang, J.; Jiang, Q.; Li, C.; Yang, J.; Su, H.; et al. 2024b. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. In *European conference on computer vision*, 38–55. Springer.

Lobo, J. M.; Jiménez-Valverde, A.; and Real, R. 2008. AUC: a misleading measure of the performance of predictive distribution models. *Global ecology and Biogeography*, 17(2): 145–151.

Mei, J.; Chen, J.; Lin, W.; Byrne, B.; and Tomalin, M. 2023. Improving hateful memes detection via learning hatefulness-aware embedding space through retrieval-guided contrastive learning. *arXiv preprint arXiv:2311.08110*.

Meng, C.; He, Y.; Song, Y.; Song, J.; Wu, J.; Zhu, J.-Y.; and Ermon, S. 2021. Sdedit: Guided image synthesis and editing with stochastic differential equations. *arXiv preprint arXiv:2108.01073*.

Mokady, R.; Hertz, A.; and Bermano, A. H. 2021. Clipcap: Clip prefix for image captioning. *arXiv preprint arXiv:2111.09734*.

Prasad, A.; Stengel-Eskin, E.; and Bansal, M. 2023. Rephrase, augment, reason: Visual grounding of questions for vision-language models. *arXiv preprint arXiv:2310.05861*.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763. PMLR.

Steck, H.; Ekanadham, C.; and Kallus, N. 2024. Is cosine-similarity of embeddings really about similarity? In *Companion Proceedings of the ACM on Web Conference 2024*, 887–890.

Su, X.; Li, Y.; Inkpen, D.; and Japkowicz, N. 2024. HateSieve: A Contrastive Learning Framework for Detecting and Segmenting Hateful Content in Multimodal Memes. *arXiv preprint arXiv:2408.05794*.

Suryawanshi, S.; Chakravarthi, B. R.; Arcan, M.; and Buitelaar, P. 2020. Multimodal meme dataset (MultiOFF) for identifying offensive content in image and text. In *Proceedings of the second workshop on trolling, aggression and cyberbullying*, 32–41.

Szegedy, C.; Vanhoucke, V.; Ioffe, S.; Shlens, J.; and Wojna, Z. 2016. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2818–2826.

Team, G.; Anil, R.; Borgeaud, S.; Alayrac, J.-B.; Yu, J.; Soricut, R.; Schalkwyk, J.; Dai, A. M.; Hauth, A.; Millican, K.; et al. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

Tschannen, M.; Gritsenko, A.; Wang, X.; Naeem, M. F.; Alabdulmohsin, I.; Parthasarathy, N.; Evans, T.; Beyer, L.; Xia, Y.; Mustafa, B.; et al. 2025. Siglip 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features. *arXiv preprint arXiv:2502.14786*.

Vinyals, O.; Toshev, A.; Bengio, S.; and Erhan, D. 2015. Show and tell: A neural image caption generator. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 3156–3164.

Wang, C.-Y.; Yeh, I.-H.; and Mark Liao, H.-Y. 2024. Yolov9: Learning what you want to learn using programmable gradient information. In *European conference on computer vision*, 1–21. Springer.

Wang, S.; Kim, D.; Taalimi, A.; Sun, C.; and Kuo, W. 2025. Learning visual grounding from generative vision and language model. In *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 8057–8067. IEEE.

Xu, K.; Ba, J.; Kiros, R.; Cho, K.; Courville, A.; Salakhudinov, R.; Zemel, R.; and Bengio, Y. 2015. Show, attend and tell: Neural image caption generation with visual attention. In *International conference on machine learning*, 2048–2057. PMLR.

Yan, S.; Bai, M.; Chen, W.; Zhou, X.; Huang, Q.; and Li, L. E. 2024. Vigor: Improving visual grounding of large vision language models with fine-grained reward modeling. In *European Conference on Computer Vision*, 37–53. Springer.

Yang, Z.; Kafle, K.; Dernoncourt, F.; and Ordonez, V. 2023. Improving visual grounding by encouraging consistent gradient-based explanations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 19165–19174.

Young, G. K. 2022. How much is too much: the difficulties of social media content moderation. *Information & Communications Technology Law*, 31(1): 1–16.

Zhang, Y.; Huang, X.; Ma, J.; Li, Z.; Luo, Z.; Xie, Y.; Qin, Y.; Luo, T.; Li, Y.; Liu, S.; et al. 2024. Recognize anything: A strong image tagging model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1724–1732.

Zhong, Y.; and Baghel, B. K. 2024. Multimodal understanding of memes with fair explanations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2007–2017.