



Article

Why So Meme? A Comparative and Explainable Analysis of Multimodal Hateful Meme Detection

Nor Saiful Azam Bin Nor Azmi , Michal Ptaszynski , Fumito Masui and Abu Nowhash Chowdhury *

Text Information Processing Laboratory, Kitami Institute of Technology, 165 Koencho, Kitami 090-0015, Hokkaido, Japan; d3237180064@std.kitami-it.ac.jp or nor.saiful.azam.nor.azmi@gmail.com (N.S.A.B.N.A.); michal@mail.kitami-it.ac.jp (M.P.); f-masui@mail.kitami-it.ac.jp (F.M.)

* Correspondence: abunowhashchy@gmail.com

Abstract

The rise of toxic content, particularly in the form of hateful memes, poses a significant challenge to social media platforms. This paper presents an empirical comparative study of unimodal and multimodal architectures for toxic content detection. Rather than proposing a novel architecture, the study evaluates the efficacy of a modular Late Fusion framework (RoBERViT) against specialized unimodal baselines (RoBERTa and ViT) and a generalist Large Multimodal (LLaVA). Both unimodal and multimodal configurations across two distinct benchmarks—the imbalanced Innopolis Hateful Memes dataset and the confounder-driven Facebook Hateful Meme dataset—were explored. Beyond quantitative metrics, this study conducts a qualitative analysis using Explainable AI (LIME) and a Large Multimodal Model (LLaVA) to investigate model reasoning. Results demonstrate that the multimodal fusion model consistently outperformed its unimodal counterparts on the Innopolis Hateful Meme dataset, achieving a toxic class F1-score of 0.6439 compared to the text-only score of 0.5794. However, on the Facebook Hateful Meme dataset, text-only models remain competitive, highlighting the “benign confounder” challenge. The qualitative analysis reveals that text remains the dominant modality, with models often relying on surface-level keywords. Notably, the Vision Transformer frequently uses text overlays as a visual proxy for hate, while the LLaVA model struggles with hallucinated toxicity in benign confounder contexts. These findings underscore the persistent challenge of achieving true multimodal understanding in hate speech detection.

Keywords: natural language processing; transformer; toxic memes; multimodal



Academic Editor: Laura Po

Received: 9 January 2026

Revised: 13 February 2026

Accepted: 18 February 2026

Published: 21 February 2026

Copyright: © 2025 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

1. Introduction

In recent years, the rapid growth of user-generated content on social media platforms has led to a significant increase in toxic and harmful material. Such content, whether textual or visual, poses widespread negative impacts on online communities and social interactions. Detecting and mitigating toxic content remains a challenging task, as it requires a comprehensive understanding of both visual and textual elements. Traditional approaches have mainly focused on either textual or visual cues; however, many toxic posts contain both text and imagery, which requires a more integrated detection system.

To address this challenge, multimodal models that combine textual and visual information have emerged as promising solutions. These models are capable of processing both modalities simultaneously, enabling them to capture subtle interactions that may be overlooked by a single-modality system. Recent advancements in deep learning, particularly

the development of Vision Transformers (ViT) for image processing and Bidirectional Encoder Representation from Transformer (BERT) for natural language understanding, have become the state-of-the-art in their respective domains. Vision Transformers demonstrate the ability to model long-range dependencies in visual data. While BERT has become a benchmark model for a wide range of natural language processing tasks, including toxic language detection [1].

This paper presents an empirical comparative study of unimodal and multimodal architectures for toxic content detection. Rather than proposing a novel architecture, we evaluate the efficacy of a modular Late Fusion framework (RoBERViT) against specialized unimodal baselines (RoBERTa and ViT) and a generalist Large Multimodal Model (LLaVA). By leveraging Explainable AI (XAI) techniques, this study aims to isolate the mechanics of multimodal fusion and diagnose the specific failure modes, such as visual noise and lexical over-reliance that occur when detecting toxic content across datasets with varying distribution.

2. Literature Review

The detection and classification of online memes, particularly those containing hateful or toxic content, has become an increasingly important research area. Memes are often used as a means for humour, commentary, or cultural expression, but they can also be exploited to spread harmful ideologies, hate speech, or misinformation. The challenge lies in the fact that memes are inherently multi-modal: they combine image, text, and often contextual or cultural cues that are not easy to decode using traditional single-modal approaches. This section outlines the evolution of research on hateful content detection, starting from early approaches in text-based analysis to recent advancements in multimodal techniques involving both language and vision transformers.

2.1. Hateful Content on the Internet

The rise of social media platforms such as Facebook, Twitter (now X), Reddit, Instagram, and YouTube has revolutionized how people communicate, share opinions, and engage in conversation. Although these platforms offer opportunities for global dialogue and community building, they also act as breeding grounds for toxic content, including hate speech, harassment, misinformation, and other forms of online abuse. Researchers have long recognized the need to moderate such content to protect users and preserve the integrity of online platforms. According to [2], hateful content generally includes speech or visual media that expresses prejudice or incites violence against individuals or groups based on characteristics such as race, religion, ethnicity, gender, nationality, sexual orientation, or disability.

Initial efforts to detect online hate focused primarily on textual content. Traditional machine learning methods such as support vector machines (SVMs), Naive Bayes classifiers, logistic regression, and decision trees were commonly used. These models typically relied on handcrafted features, including bag-of-words representations, term frequency-inverse document frequency (TF-IDF), and syntactic or part-of-speech features extracted from text [3,4]. While effective to some extent, these approaches were often limited in their ability to handle subtle, implicit, or coded hate speech, especially when context and cultural understanding were required. Furthermore, keyword-based systems frequently produced false positives and failed to generalize well across different domains or languages.

As hateful content on the Internet became more diverse, researchers began to recognize the need to detect hate in more complex and multimodal formats, such as memes. Memes can embed offensive messages that are highly dependent on context, irony, or visual-textual interplay. This realization shifted the focus toward more advanced models capable of integrating multiple types of data.

2.2. Automatic Hateful Content Detection

The field of automatic hateful content detection has progressed rapidly with the advancement of deep learning, especially in natural language processing and computer vision. With the growing availability of large datasets and pretrained models, researchers have begun to develop systems that automatically detect toxic content in both text and images, often using a combination of the two for better accuracy.

Recent studies have highlighted the limitations of unimodal approaches, those that use either text or image alone for analyzing memes or other forms of multimodal content. For example, the image component of a meme may appear benign in isolation, while the accompanying text adds a layer of offensive meaning, or vice versa [5]. This has led to a surge in interest in multimodal systems that analyze both modalities simultaneously.

One of the most influential works in this area is the Facebook Hateful Memes Challenge introduced by [6]. The dataset consists of over 10,000 memes labelled for hatefulness, and includes both image and text components. The challenge spurred numerous studies aimed at building more robust classifiers capable of detecting hateful memes using multimodal techniques.

2.2.1. Classic Machine Learning and Transformers

Machine learning approaches to hateful content detection have evolved significantly over time. Early methods relied on handcrafted linguistic features and traditional classifiers, while modern techniques incorporate deep learning architectures such as transformers, which can model complex dependencies within textual and multimodal data. The rise of transformers, particularly in text classification tasks, has revolutionized hateful content detection [7].

2.2.2. Text Classification Transformers

In the area of text classification, transformer-based models have become state-of-the-art due to their powerful ability to model complex relationships within language. The transformer architecture, introduced by [8], uses self-attention mechanisms to weigh the importance of different words in a sequence, allowing for a more nuanced understanding of context compared to earlier RNN or CNN-based approaches.

Among the many transformer models available today, RoBERTa [9] stands out as one of the most effective for hate speech detection. RoBERTa is a robustly optimized version of BERT that improves performance through better pretraining, larger batch sizes, and more training data. When fine-tuned on domain-specific datasets, RoBERTa has shown impressive results in identifying hate speech, offensive language, and toxic comments across various social media platforms [10,11]. Because of its ability to understand subtle linguistic cues such as sarcasm, metaphor, and context-switching, RoBERTa is particularly useful in cases where hate is conveyed in indirect or ironic ways.

2.2.3. Image Detection Using Transformer Models

In the field of image analysis, Vision Transformers (ViT) have revolutionized the way models process visual information. Traditionally, Convolutional Neural Networks (CNNs) dominated computer vision tasks, using localized filters to detect patterns in image regions. However, CNNs often struggle to capture long-range dependencies or contextual relationships across an image, which are crucial for understanding complex content such as memes.

The Vision Transformer, introduced by [12], adapts the transformer architecture for image data by splitting an image into fixed-size patches, embedding each patch, and feeding the sequence into a transformer encoder. Unlike CNNs, ViTs use global self-attention to model relationships across all parts of an image simultaneously, making them better suited for tasks where spatial context matters.

ViTs have demonstrated strong performance in image classification benchmarks like ImageNet and have shown potential in detecting harmful content in memes, where understanding the global composition of an image is important. Moreover, they are data-efficient when fine-tuned on smaller datasets, especially with techniques like distillation or transfer learning [13].

2.3. Meme Detection and Classification

Detecting memes involves challenges that are not usually present in standard image or text classification tasks. A meme's meaning is often derived from the interaction between image and text, which may rely on shared cultural knowledge, political context, or internet slang. Early research used CNNs for image processing and separate NLP models like LSTMs or RNNs for the text component [5,6]. These models would then combine the two outputs using simple fusion techniques like concatenation or averaging.

However, these early multimodal models lacked the ability to capture deeper semantic interactions between text and image. They often treated the two modalities as independent, missing out on the rich contextual interplay that makes memes so nuanced.

2.4. Multimodal Transformers

To address this gap, researchers began exploring multimodal transformers that process both image and text within a unified architecture. One of the most notable models in this space is CLIP (Contrastive Language-Image Pretraining), developed by [14]. CLIP learns a joint embedding space for text and images, allowing it to understand how certain captions relate to visual concepts. When applied to meme classification, CLIP has shown promising results by aligning visual and linguistic cues in a way that enables the detection of subtle forms of hate or offensive humor.

Another successful approach combines CLIP with language models like BERTweet [15], which is pretrained specifically on Twitter data. This integration helps handle the informal, abbreviated, and slang-filled language often found in meme captions or tweets. The use of BERTweet improves the model's sensitivity to social media-specific language patterns, further boosting meme classification accuracy.

Some researchers have also explored interpretable approaches, such as Explainable Deep Neural Networks (xDNNs), which use prototype-based learning to classify images while offering human-readable explanations [16]. This interpretability is especially valuable in content moderation scenarios, where human reviewers need to understand why a model flagged a piece of content.

Other promising multimodal frameworks include VisualBERT [17], ViLBERT [18], and MMBT [19], each designed to fuse visual and textual embeddings more effectively. These models have been applied to meme detection tasks with improved performance over traditional methods.

2.5. Large Multimodal Models

Large Language Models (LLMs) like GPT-3 [20] and GPT-4 have significantly advanced the field of natural language understanding by enabling zero-shot and few-shot learning across a wide range of tasks. These models are trained on massive text corpora and can generate highly contextual, human-like language. In the context of toxic content detection, LLMs offer the advantage of understanding implicit and nuanced language, including sarcasm, coded hate, or cultural references that traditional models often miss.

However, in the context of meme detection, text-only LLMs are limited by their inability to directly process visual information. Early adaptations addressed this by employing a "pipeline" approaches, where visual content was converted into textual description (captions) which were often fed into models like LLaMA [21] or PaLM [22] for classification.

While effective for explicit content, these pipeline methods often suffer from information loss, as subtle visual cues or text-image irony are lost during the captioning process.

To overcome these limitations, the field has shifted towards Large Multimodal Models (LMMs) such as GPT4-V [23] and LLaVA [24]. Unlike pipeline approach, LMMs process visual and textual tokens in a shared embedding space, allowing for native multimodal reasoning. LLaVA, for instance, utilizes instruction tuning to align a vision encoder (CLIP) with a Language decoder (Vicuna), enabling it to not only classify images but also generate coherent explanations for its decisions [24]. This generative capability offers a significant advantage for toxic content detection, as it allows the model to articulate “why” a meme is hateful, providing a layer of interpretability that traditional discriminative models (such as BERT or ResNet) lack. However, despite their reasoning capabilities, LMMs can be computationally expensive and prone to “hallucination”, where the model generates plausible but factually incorrect details about the image content.

3. Methodology

This section details the experimental setup, including dataset selection, model configurations, evaluation metrics, and experimental procedures used to assess the performance of Vision Transformers (ViT) and RoBERTa in multimodal hateful meme detection.

3.1. Dataset

The experiments are conducted using the Innopolis Hateful Meme dataset. This dataset offers a collection of multimodal samples that include both image and text elements, which will be an appropriate and challenging benchmark for assessing the performance of models designed to detect hateful or toxic content in memes.

3.1.1. Facebook Hateful Meme Dataset

The Facebook Hateful Meme dataset, introduced by [6], is an important benchmark utilized for evaluating multimodal hate speech detection system. This dataset consists of over 10,000 memes that are binary labelled as either hateful or non-hateful.

A key characteristic of this dataset is its construction, which includes “benign confounders” designed to prevent unimodal reliance. By pairing benign images with toxic text or vice versa, the dataset ensures that the hateful sentiment is derived specifically from the interaction between the visual and textual modalities, rather than from either modality in isolation [5,6]. This structure necessitates the models that can perform true multimodal reasoning to distinguish between hateful and benign content effectively.

As detailed in Table 1, the dataset employs a controlled distribution strategy. The Validation (Dev) set strictly balanced (50:50) to ensure unbiased hyperparameter tuning and model selection. The Training set maintains a ratio of approximately 36% toxic samples (3050 toxic vs. 5450 non-toxic), providing sufficient minority-class examples without artificial over-representation. The Test set follows a similar distribution (approximately 34% toxic), testing the model’s ability to generalize to unseen, slightly imbalanced data distributions.

Table 1. Distribution of Toxic and Non-Toxic samples across datasets.

	Innopolis Hateful Memes		Facebook Hateful Memes	
	Non-Toxic	Toxic	Non-Toxic	Toxic
Train	14,195	2054	5450	3050
Dev	1775	256	250	250
Test	1775	257	662	338

The Facebook Hateful Meme dataset used in this study is divided into five subsets:

- Training set: Used for model training
- Validation (seen) set: Used for hyperparameter tuning and intermediate evaluation on memes with characteristics similar to the training set
- Validation (unseen) set: Used to evaluate the model's generalization to new meme formats not seen during training
- Test (seen) set: Used for final model assessment on familiar meme types
- Test (unseen) set: Used for final model assessment on completely unseen meme formats

The Facebook Hateful Meme dataset employs a controlled distribution strategy. While the Validation set is strictly balanced (50:50) to ensure unbiased model selection, the Training set maintains a moderate imbalance (approximately 36% toxic). However, this distribution still provides a significantly higher density of toxic samples compared to the Innopolis Hateful Meme dataset, providing a rigorous baseline for assessing the architecture's effectiveness in a controlled environment without the extreme scarcity found in "in-the-wild" data.

3.1.2. Innopolis Hateful Meme Dataset

The Innopolis Hateful Meme dataset is used in this study to provide a more diverse and challenging set of hateful and benign memes. It is specifically curated to address the limitation of existing datasets by incorporating a broader range of cultural and contextual variation in hate speech [25].

The Innopolis Hateful Meme dataset comprises multimodal memes in which the interaction between text and image determines the label. The dataset is divided into three subsets.

- Training set: Used for model training
- Validation set: Used for hyperparameter tuning and intermediate evaluation
- Test set: A held-out portion used for final model assessment

The dataset is constructed to reduce biases found in previous meme datasets, thereby enabling a more balanced and comprehensive evaluation of multimodal classification models. It supports a robust assessment of the effectiveness of Vision Transformer (ViT) and RoBERTa-based models in detecting hateful content across varied multimodal contexts.

The Innopolis Hateful Meme dataset serves as a critical benchmark for evaluating multimodal classifiers and provides a realistic foundation for developing models aimed at detecting hate speech in online environments.

3.2. Experimental Procedure

Selection of Models

While the landscape of transformer architecture continues to evolve with models such as DeBERTa and Swin Transformer, this study deliberately utilized RoBERTa and the standard Vision Transformer (ViT) as foundation baselines. This selection is driven by the need for controlled experimental environment:

- Architectural Transparency: RoBERTa and ViT represent the canonical architectures for their respective modalities. Utilizing these standard-bearers allows for the isolation of the effects of the fusion strategy in this case Late Fusion and the dataset characteristics. Without the confounding variables introduced by the complex attention mechanisms found in newer architectures.
- Explainability Focus: The primary contribution of this work is an explainable analysis of multimodal failure modes. Standard architectures allow for a more direct application of interpretability tools like LIME, facilitating a clearer diagnosis of "visual noise" and lexical reliance.

Future iteration of this work will expand this comparative framework to include next-generation backbones to assess if advanced architectural inductive biases can mitigate the fusion limitation identified in this study.

To assess the effectiveness of multimodal approaches for toxic and hateful meme detection, this study employs a comprehensive experimental framework that integrates both image and text modalities. The experimental design centers on evaluating Vision Transformer for image-based representation learning and RoBERTa for textual feature extraction. Both unimodal and multimodal configurations are explored to understand the individual and combined strength of visual and textual features in detecting toxicity within memes.

The experimental pipeline is applied to two distinct benchmarks: The Innopolis Hateful Meme Dataset and the Facebook Hateful Meme dataset. This dual-dataset approach allows for a robust evaluation by testing the models against two distinct toxicity distribution: The Innopolis Hateful meme dataset, which tests performance on “in-the-wild” imbalanced data with explicit hate symbols, and the Facebook Hateful meme dataset, which tests reasoning capabilities against “benign confounders” and imbalanced scenario and the implicit, balanced scenario. These datasets contain meme samples that pair an image with associated textual content, such as caption or phrases.

Prior to the final evaluation, preliminary tuning runs were conducted to empirically determine the optimal hyperparameters. Learning rates in the range of (1×10^{-4} , 3×10^{-5} , 1×10^{-5}) and varying batch sizes constrained by GPU memory were explored. The reported configuration (Learning Rate 1×10^{-5} , 20 epochs) was selected as it yielded the highest stability and F1-score on the validation set during these initial sweeps.

The evaluation begins with Vision Transformer (ViT) models. ViT, an architecture based entirely on a self-attention mechanism, is employed to extract visual features from meme images. Two variants of Vision transformer are adopted: ViT-Base with a patch size of 16 and input resolution of 224×224 , and ViT-Large with a patch size of 16 and input resolution of 384×384 . Both models are fine-tuned on the target dataset to adapt their visual representations to the specific subtleties of memes, which often include stylized fonts, layered text, and diverse background textures. Figure 1 shows the flowchart of ViT processes. Image preprocessing involves resizing, normalizing, and random augmentation to enhance generalization. The performance of ViT models trained with and without image augmentation is compared to assess the effect of data augmentation on visual learning.

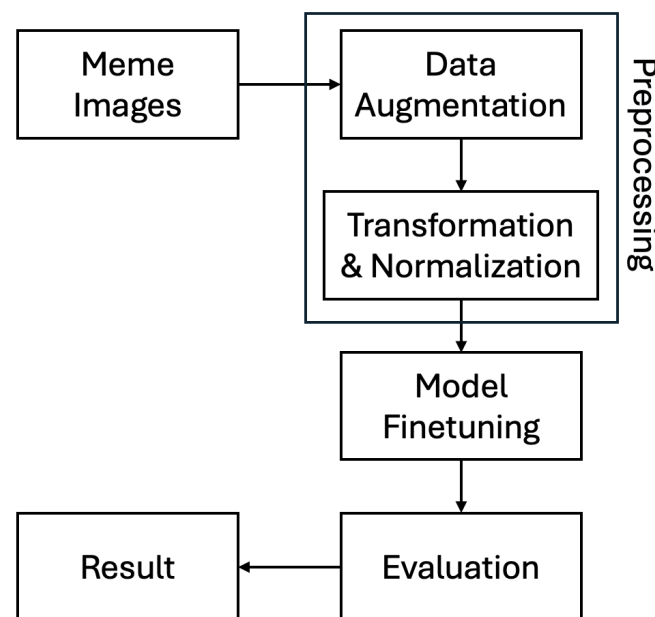


Figure 1. Vision Transformer Configuration.

For textual analysis, Figure 2 shows a flowchart where RoBERTa is employed, a robustly optimized transformer-based language model pretrained on large-scale corpora.

Both the base and large variants of RoBERTa are fine-tuned to capture textual cues indicative of toxicity. The text used in the training is sourced from the annotations provided in the dataset. Preprocessing includes tokenization and normalization aligned with RoBERTa's input requirements. Fine-tuning is applied to the entire model to enable contextual adaptation to meme-specific language, which often includes sarcasm, slang and coded hate speech.

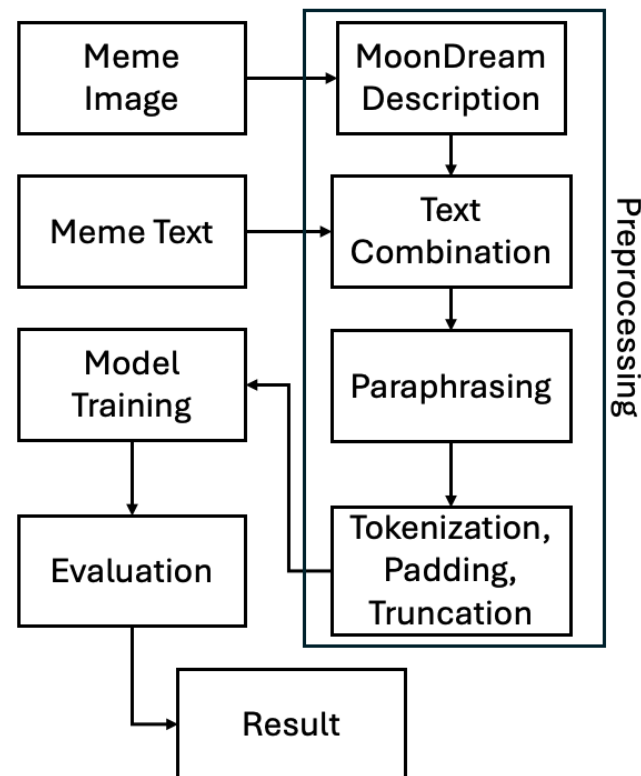


Figure 2. RoBERTa Configuration.

To further enhance the training data, two data augmentation strategies are employed. First, additional textual descriptions are generated using MoonDream, a vision-to-text model that describes image content. Second, T-5 based paraphrasing is used to expand the dataset with semantically equivalent variants of the descriptions. Three configurations are compared: the original non-augmented dataset, the dataset with image descriptions, and the dataset with both description and paraphrased variants.

In the multimodal configuration shown in Figure 3, a late fusion strategy is implemented to integrate the visual and textual modalities. The Vision Transformer processes the image input, while RoBERTa encodes the textual input, including both the original meme text and the generated descriptions in the Figure 2. The final hidden states from both modalities, specifically the [CLS] token representation from ViT and RoBERTa, are concatenated to form a unified multimodal feature vector. This combined vector is passed through a fully connected classification head consisting of two dense layers with a ReLU activation function and dropout regularization to mitigate overfitting. A softmax function is applied at the output layer to perform binary classification, determining whether a meme is toxic or non-toxic, or hateful versus benign, depending on the target label in the dataset.

To further clarify the multimodal fusion process relative to the unimodal baselines, a detailed architectural diagram visualizing the two parallel feature extraction streams merging via a late-fusion strategy has been constructed.

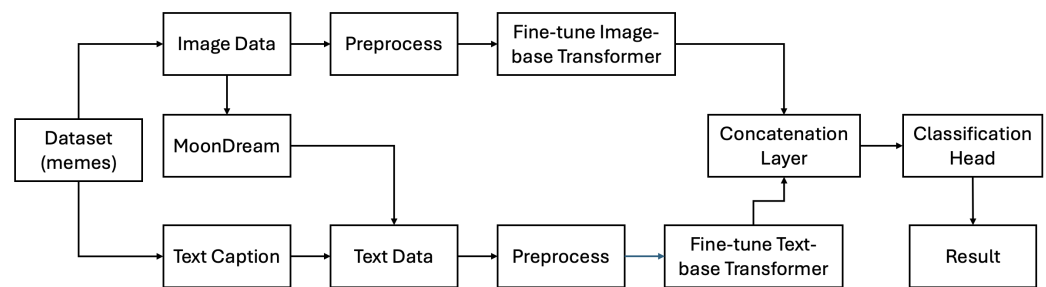


Figure 3. Multimodal Configuration.

Figure 4 illustrates this proposed modular architecture. Visual features extracted by the Vision Transformer (ViT) and textual features extracted by RoBERTa are concatenated (Late Fusion) before passing through a dense classification head. The pipeline operates as follows:

1. **Visual Stream:** The raw image is processed by the Vision Transformer (ViT), extracting a 768-dimensional feature vector from [CLS] token. For the Base model, this vector is 768-dimensional, and for the Large model used in the experiments is 1024-dimensional.
2. **Textual Stream:** The meme text is augmented with a caption generated by the MoonDream model. This combined text is tokenized and processed by RoBERTa, yielding a contextual embedding (768-d for Base, 1024-d for Large).
3. **Fusion & Classification:** These two vectors are concatenated to form a unified multimodal representation (1536-d as shown). This combined vector is passed through a classification head explicitly designed to reduced dimensionality while preventing overfitting. It consists of a first Dense Layer with ReLU activation (reducing 1536 dimension to 512), followed by a Dropout layer ($p = 0.1$), and finally a Linear layer (reducing 512 dimensions to 2) to predict the final toxicity probability via Softmax.

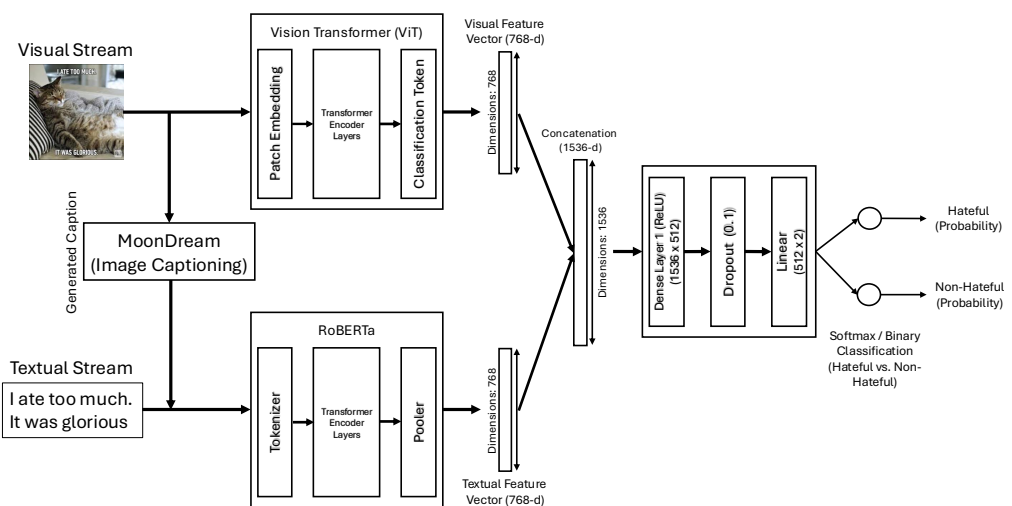


Figure 4. Detailed Visualization of Multimodal Configuration.

This modular design allows us to isolate the contribution of each modality by comparing the performance of the fused model against the individual streams.

For evaluating the model effectiveness, standard classification metrics are reported, including accuracy, precision, recall, and F1-Score. Each configuration is executed multiple times with different random seeds, and the average results are reported to ensure reproducibility and minimize the impact of random variation. Comparisons are conducted across the fine-tuning scenarios to demonstrate the adaptability and the robustness of both unimodal and multimodal approaches.

3.3. Experimental Setup

The RoBERTa experiments are conducted using the PyTorch deep learning framework in a Python environment, running on a machine equipped with four NVIDIA 2080TI Graphics Processing Units, 135 GB of system memory, and an Intel Core i9-9980XE Central Processing Unit. The unimodal models are trained with a batch size of 36 for the base model and 8 for the large model. Training is performed for up to 20 epochs, utilizing a Cosine Annealing learning rate scheduler to dynamically adjust the learning rate throughout training. This scheduling strategy is to smooth the rapid convergence during the early epochs and gradually reduce the learning rate in later stages, leading to more stable updates and improved generalization.

The Vision Transformer (ViT) experiments are also implemented using the PyTorch framework (version 2.6.0) in a Python environment. These experiments are conducted on a machine equipped with an NVIDIA RTX A6000 Graphics Processing Unit, 135 GB of system memory, and an Intel Core i9-7920X Compute Processing Unit. The base ViT Model with a patch of 16 and resolution of 224×224 is trained with a batch size of 32, while the large ViT model with a patch size of 16 and resolution of 384×384 is trained with a batch size of 16. Training is carried out for up to 20 epochs with a learning rate of 1×10^{-5} . To enhance model generation, a set of image augmentation techniques is applied during training, including random horizontal flipping, random rotation, color jittering, and random resized cropping.

The multimodal RoBERTa + Vision Transformer experiment, referred to as RoBERViT, is also conducted using the PyTorch framework on the same machine used for the Vision Transformer experiments, chosen for its superior computational capabilities. The multimodal model is trained with a batch size of 32 for the base configuration and 16 for the large configuration. All preprocessing and implementation strategies applied in the unimodal experiments, including fine-tuning protocols, learning rate scheduling, and data augmentation, are also adopted in the multimodal setup, based on their observed effectiveness in the individual experiments.

The result and performance evaluation of these experiments are discussed in Section 4, starting with RoBERTa experiments, followed by the Vision Transformer results and lastly the multimodal RoBERViT evaluation.

3.4. Qualitative Analysis Methodology

To complement the quantitative performance metrics, a qualitative analysis is conducted to interpret the reasoning process of the models. Two distinct Explainable AI (XAI) approaches are employed, tailored to the specific architecture of the models.

3.4.1. Local Interpretable Model-Agnostic Explanation (LIME)

For the discriminative models (RoBERTa and Vision Transformer), the Local Interpretable Model-agnostic Explanation (LIME) framework [26] is utilized. LIME approximates the complex, non-linear decision boundaries of deep neural networks with a locally interpretable linear model. This is achieved by perturbing the input data (for example: masking words in text or superpixels in images) and observing the resulting changes in prediction confidence.

- Textual LIME: For RoBERTa, LIME identifies specific tokens that make the strongest positive or negative contribution to the “Hateful” classification.
- Visual LIME: For the Vision Transformer, LIME segments the image into superpixels. It then highlights the specific visual regions that the model attends to when making a prediction, revealing whether the model focuses on semantic object (e.g., a person face) or background noise.

3.4.2. Generative Reasoning (LLaVA)

As LLaVA (specifically LLaVA-v.1.5-7B) is a generative vision-language model, standard perturbation-based XAI methods are less suitable. Instead, a “Self-Explanation” prompting strategy is adopted. A structured two-stage prompting strategy was implemented to extract both qualitative reasoning and quantitative confidence scores. The model was prompted to not only classify the meme but to explicitly articulate its reasoning. This process involves a structured prompt requiring the model to:

1. Analyze the textual content for the toxic language.
2. Analyze the visual content for hateful symbols or context.
3. Explain how the interaction between text and image leads to the final classification.

This approach allows for the detection of “hallucination” (where the model describes visual elements that are not present) and assesses the model’s ability to perform true multimodal grounding.

3.4.3. Qualitative Explanation Generation

To explore how the LLaVA model arrives at its prediction, a structured prompting strategy was employed to get a qualitative explanation. The model was provided with a multimodal input (image and text) and guided through a set of targeted instructions. The prompt consisted of three components:

“Based on the examples, analyze the image. The text in the image is: {text}”

- Classification: Is this meme hateful? Answer with only ‘Yes’ or ‘No’.
- Textual Analysis: Explain your reasoning. Which specific words or phrases contribute most to your decision, and why?
- Visual Analysis: Was the image a factor in your decision? Answer ‘Yes’ or ‘No’. If yes, describe the specific visual elements that were most influential and explain how they interact with the text to create the meme’s meaning.

3.4.4. Quantitative Confidence Score Extraction

While Qualitative Explanation Generation emphasized interpretability, quantitative confidence score extraction was designed to derive a measurable confidence score from the model. To achieve this, the model was presented with a simplified and highly constrained classification prompt “logit_prompt” that only required a binary answer (Yes or No). A direct forward pass was then performed to extract the raw, unnormalized logit values for these two target tokens from the models final prediction layers. These logits, representing the models internal confidence prior to normalization, were converted into a probability distribution using softmax function.

These process combined qualitative interpretability with quantitative precision, enabling a more comprehensive evaluation of LLaVA’s decision-making process in multimodal hateful meme classification.

4. Experiments and Results

This section presents the experiments conducted, and the results obtained for three models: RoBERTa (text-only), Vision Transformer (image-only), and the combined RoBERTa + Vision Transformer model, referred to as RoBERViT. Each experiment is described in terms of dataset preparation, training procedure, and evaluation result. Detailed observations are provided following each experiment to highlight the strengths, limitations and the improvements introduced by different training and augmentation strategies.

4.1. RoBERTa Experiments

The RoBERTa experiments focus on the textual modality of the memes. To ensure a comprehensive evaluation, the model was fine-tuned on two distinct benchmarks: the Facebook Hateful meme dataset and the Innopolis Hateful Meme dataset. The following subsection details the performance across progressive optimization stages for each dataset.

4.1.1. RoBERTa on Facebook Hateful Meme Dataset

The Facebook Hateful meme dataset is utilized to evaluate the model's performance on a widely recognized benchmark with a distinct distribution and structure compared to the Innopolis Hateful Meme Dataset. As described in the dataset section, this collection comprises multimodal entries pre-divided into training, validation (seen/unseen), and test (seen/unseen) sets [6]. Unlike the custom splits required for Innopolis Hateful Meme, this experiment adheres to the partition provided by the challenge to ensure comparability with existing literature.

To maintain consistency with the methodology applied to the Innopolis Hateful Meme dataset, the textual data was enriched using identical augmentation strategies. The MoonDream1 model was employed to generate visual descriptions for every image in the Facebook Hateful Meme dataset, which were then appended to the original meme text. Furthermore, the T-5 based paraphrasing model was utilized to create semantically equivalent variation of the captions. This approach ensures that the RoBERTa model is fine-tuned on a semantically dense and linguistically diverse corpus, mirroring the rigorous preprocessing pipeline established in the subsequent experiment.

The following subsections detail the performance of the RoBERTa model across the progressive optimization stages, ranging from the baseline configuration to fully augmented definitions.

Baseline Results

The initial experiments involved fine-tuning RoBERTa on the original textual content of the Facebook Hateful memes training set without additional augmentation. A learning rate of 1×10^{-5} was maintained, with training duration scaling according to the model size. The result on the test set are summarized in the Table 2.

Table 2. RoBERTa-Finetuned on Facebook Hateful Meme Dataset Result.

Base Model	Precision	Recall	F1	Support
Non-toxic	0.5849	0.7902	0.6722	510
Toxic	0.6559	0.4163	0.5094	490
Accuracy			0.607	1000
macro avg	0.6204	0.6033	0.5908	1000
weighted avg	0.6197	0.607	0.5924	1000
Large Model	Precision	Recall	F1	Support
Non-toxic	0.5849	0.7902	0.6722	510
Toxic	0.6559	0.4163	0.5094	490
Accuracy			0.6070	1000
macro avg	0.6204	0.6033	0.5908	1000
weighted avg	0.6197	0.6070	0.5924	1000

The results in the Table 2, show a notable inconsistency between precision and recall, particularly for the toxic class. While the model achieves reasonable accuracy, the F1-scores around 0.51 for toxic class indicate significant room for improvement in correctly identifying hateful content without generating excessive false positives.

Oversampling Strategy

To address the severe class imbalance (where non-toxic samples heavily outnumber toxic ones), random oversampling was applied to the minority “Toxic” class. This involved duplicating toxic samples in the training batch until the class distribution was approximately equal. This strategy aims to improve sensitivity (Recall), ensuring the model is penalized equally for missing a toxic meme as it is for misclassifying a benign one, thereby preventing the classifier from defaulting to the majority “Non-Toxic” label. The result of these experiments are presented in Table 3.

Table 3. RoBERTa-Finetuned with 3× Oversampling on Facebook Hateful Meme Dataset Result.

Base Model	Precision	Recall	F1	Support
Non-toxic	0.6282	0.6196	0.6239	510
Toxic	0.6057	0.6184	0.614	490
Accuracy			0.6190	1000
macro avg	0.6189	0.6190	0.6189	1000
weighted avg	0.6191	0.6190	0.6190	1000
Large Model	Precision	Recall	F1	Support
Non-toxic	0.6358	0.6059	0.6205	510
Toxic	0.6089	0.6388	0.6235	490
Accuracy			0.622	1000
macro avg	0.6224	0.6223	0.6220	1000
weighted avg	0.6226	0.6220	0.6220	1000

As shown in Table 3, oversampling led to a more balanced performance between precision and recall for both classes. Specifically, the toxic class recall improved significantly compare to the baseline, suggesting that the model became less biased toward the non-toxic class.

Learning Rate Optimization

To further stabilize training and potentially improve generalization, a cosine annealing learning rate scheduler was introduced. The performance results are reported in Table 4.

Table 4. RoBERTa-Finetuned with Learning Rate Optimization and 3× Oversampling on Facebook Hateful Meme Dataset Result.

Base Model	Precision	Recall	F1	Support
Non-toxic	0.6809	0.8157	0.7422	510
Toxic	0.7584	0.6020	0.6712	490
Accuracy			0.7110	1000
macro avg	0.7196	0.7089	0.7067	1000
weighted avg	0.7188	0.7110	0.7074	1000
Large Model	Precision	Recall	F1	Support
Non-toxic	0.6936	0.8078	0.7464	510
Toxic	0.7586	0.6286	0.6875	490
Accuracy			0.7200	1000
macro avg	0.7261	0.7182	0.7169	1000
weighted avg	0.7255	0.7200	0.7175	1000

Based on the Table 4, the introduction of the learning rate scheduler resulted in a performance boost for the Base model, achieving an accuracy of 0.7110. However, the Large model did not show the same level of improvement, indicating that larger architecture may require different hyperparameter tuning strategies on this specific dataset.

Incorporating Image Description

Previous experiments suggested that the performance is constrained by the lack of explicit semantic cues in the text alone. To address this, the MoonDream Image caption model was used to generate short description for each meme image, providing an additional modality to bridge the semantic gap. A new version of the dataset was created by appending these descriptions as an additional text column.

MoonDream-Enhanced Data

To further enrich the textual modality, descriptions generated by the MoonDream model were integrated into the training data. The result are presented in Table 5.

Table 5. RoBERTa-Finetuned with MoonDream Description, Learning Rate Optimization and 3× Oversampling on Facebook Hateful Meme Dataset Result.

Base Model	Precision	Recall	F1	Support
Non-toxic	0.6809	0.8157	0.7422	510
Toxic	0.7584	0.6020	0.6712	490
Accuracy			0.7110	1000
macro avg	0.7196	0.7089	0.7067	1000
weighted avg	0.7188	0.7110	0.7074	1000
Large Model	Precision	Recall	F1	Support
Non-toxic	0.6936	0.8078	0.7464	510
Toxic	0.7586	0.6286	0.6875	490
Accuracy			0.7200	1000
macro avg	0.7261	0.7182	0.7169	1000
weighted avg	0.7255	0.7200	0.7175	1000

In Table 5, the inclusion of MoonDream description provided a performance increase for the Large model pushing the accuracy to 0.7200 and improving the toxic F1-score to 0.6875. This suggests that the generated caption successfully provide explicit textual proxies for visual elements that were previously inaccessible to the text-only model.

T5 Paraphrasing Implementation.

Finally, paraphrases were generated using the T5 model and added to the original caption and MoonDream descriptions. This was intended to increase textual diversity. The results are shown in the Table 6.

Table 6. RoBERTa-Finetuned with MoonDream Description, Paraphrased, Learning Rate Optimization and 3× Oversampling on Facebook Hateful Meme Dataset Result.

Base Model	Precision	Recall	F1	Support
Non-toxic	0.6425	0.4510	0.5300	510
Toxic	0.5639	0.7388	0.6396	490
Accuracy			0.5920	1000
macro avg	0.6032	0.5949	0.5848	1000
weighted avg	0.6039	0.5920	0.5837	1000
Large Model	Precision	Recall	F1	Support
Non-toxic	0.6475	0.4647	0.5411	510
Toxic	0.5694	0.7367	0.6423	490
Accuracy			0.5980	1000
macro avg	0.6085	0.6007	0.5917	1000
weighted avg	0.6093	0.5980	0.5907	1000

Similar to the Innopolis results (discussed in the next section), the paraphrasing strategy led to a significant degradation in performance, with accuracy dropping below 60%. This reinforces the finding that extensive paraphrasing may dilute the specific linguistic triggers required for accurate hate speech detection in memes.

4.1.2. RoBERTa on Innopolis Hateful Memes

The Innopolis Hateful Meme dataset comprises four key components: image ID, image file path, associated textual content, and binary hatefulness label. The dataset contains a total of 20,312 memes, of which 2567 are labelled as toxic and 17,745 as benign. While the term “benign” may initially appear ambiguous, the dataset documentation clearly defines it as non-toxic [25].

As the original dataset is not pre-divided into training, development, and testing subsets, a custom split was conducted prior to model training. The dataset was partitioned into training (80%), development (10%), and test (10%), with class-balanced sampling to ensure equal representation of hateful and benign labels across all splits. This approach helps prevent training bias and supports more reliable evaluation.

Given the relatively small number of toxic samples, additional syntactic and semantic information was introduced to enrich the training data. To this end, the image captioning model MoonDream1 (<https://huggingface.co/vikhyatk/mondream1> (accessed on 10 December 2025)) created by a user vikhyatk from Hugging Face (<https://huggingface.co/> (accessed on 10 December 2025)) was used to generate descriptive textual annotation for each image. These generated descriptions were appended to the corresponding meme text in all dataset splits (training, development, and test), effectively enhancing the semantic richness of each sample.

To further increase textual diversity and dataset size, a paraphrasing strategy was employed using a T5 based paraphrasing model. This stage allows the generation of semantically equivalent variants for both the original meme caption and the descriptive text generated by MoonDream. For example, if MoonDream described an image as “A cat sitting on a table”, the T5 paraphraser might generate variants such as “A feline resting atop a desk”. This increases the linguistic diversity of the training data, preventing the model from overfitting to specific syntactic patterns associated with the image descriptions. As a result, the dataset was synthetically expanded, improving the language model’s robustness during fine-tuning and increasing its exposure to varied linguistic pattern.

Following the data augmentation process, the RoBERTa model was fine-tuned on the enhanced dataset. A cosine annealing learning rate scheduler was applied to facilitate efficient convergence and reduce overfitting. The model was trained for 20 epochs, with hyperparameters tuned based on performance metrics evaluated on the development set.

Baseline Results

Initial experiments involved fine-tuning RoBERTa on the original dataset without augmentation or oversampling. A learning rate of 1×10^{-5} was used. Training required approximately 1 h and 40 min for the base model and 9 h and 20 min for the large model. The results are summarized in Table 7.

The results demonstrate strong performance on the non-toxic class, with F1-scores exceeding 0.93 for both model sizes. However, the models underperform on the toxic class with F1-scores of 0.527 for the base model and 0.551 for the large model. This inconsistency highlights the challenges posed by the class imbalance, where toxic examples are fewer than non-toxic ones.

Table 7. RoBERTa-Finetuned Result.

Base Model	Precision	Recall	F1	Support
Non-toxic	0.9281	0.9448	0.9363	1775
Toxic	0.5644	0.4942	0.527	257
Accuracy			0.8870	2032
macro avg	0.7463	0.7195	0.7317	2032
weighted avg	0.8821	0.8878	0.8846	2032
Large Model	Precision	Recall	F1	Support
Non-toxic	0.9326	0.9431	0.9378	1775
Toxic	0.5738	0.5292	0.5506	257
Accuracy			0.8907	2032
macro avg	0.7532	0.7361	0.7442	2032
weighted avg	0.8872	0.8907	0.8888	2032

Oversampling Strategy

To address the issue of class imbalance in the Innopolis Hateful Meme dataset, random oversampling was applied. The results of these experiments are presented in Table 8.

As shown in Table 8, oversampling the minority toxic class led to a notable improvement in recall for the base RoBERTa model from 0.494 to 0.603, which subsequently increased its F1 scores. This indicates that oversampling successfully enhanced the model's ability to detect toxic content. However, for the large RoBERTa model, only marginal performance gains were observed. The relatively low F1 scores for the toxic class, despite oversampling, suggest that larger models may require more subtle augmentation strategies. These models, due to their greater capacity, may be more susceptible to overfitting when exposed to duplicated minority-class examples.

To address this limitation, subsequent experiments explored learning rate adjustment and more sophisticated augmentation techniques aimed at further improving generalization while avoiding overfitting. These refinements are discussed in the next section.

Table 8. RoBERTa-Finetuned with 3× Oversampling Result.

Base Model	Precision	Recall	F1	Support
Non-toxic	0.9410	0.9172	0.9290	1775
Toxic	0.5132	0.6031	0.5546	257
Accuracy			0.8775	2032
macro avg	0.7271	0.7601	0.7418	2032
weighted avg	0.8869	0.8775	0.8816	2032
Large Model	Precision	Recall	F1	Support
Non-toxic	0.9300	0.9425	0.9362	1775
Toxic	0.5622	0.5097	0.5347	257
Accuracy			0.8878	2032
macro avg	0.7461	0.7261	0.7354	2032
weighted avg	0.8835	0.8878	0.8854	2032

Learning Rate Optimization

To further stabilize training and potentially improve generalization, a cosine annealing learning rate scheduler was introduced. The performance results after applying this scheduler, combined with toxic oversampling techniques, are reported in Table 9.

As shown in Table 9, the base RoBERTa model maintained comparable performance to the previous experiment, with a slight improvement in the recall of the toxic class. This suggests that the cosine scheduler provided a more stable learning trajectory, especially

when combined with oversampling. However, the large RoBERTa model exhibited a slight degradation in performance, particularly in toxic class recall and F1 score. This indicates that the large model may require more specialized hyperparameter tuning beyond the current configuration to realize its full potential.

Table 9. RoBERTa-Finetuned with Learning Rate Optimization and $3 \times$ Oversampling Result.

Base Model	Precision	Recall	F1	Support
Non-toxic	0.9369	0.9285	0.9327	1775
Toxic	0.5348	0.5681	0.5509	257
Accuracy			0.8829	2032
macro avg	0.7358	0.7583	0.7418	2032
weighted avg	0.8860	0.8829	0.8844	2032
Large Model	Precision	Recall	F1	Support
Non-toxic	0.9244	0.9363	0.9303	1775
Toxic	0.5171	0.4708	0.4929	257
Accuracy			0.8775	2032
macro avg	0.7207	0.7036	0.7116	2032
weighted avg	0.8729	0.8775	0.8750	2032

Incorporating Image Description

Previous experiments confirmed that while the base RoBERTa model could moderately benefit from oversampling and learning rate scheduling, performance on toxic memes remained constrained, largely due to the lack of explicit semantic cues in the memes' text or image alone. In many memes, toxic meaning emerges not from the caption or image in isolation, but from their interplay. To address this, the MoonDream image captioning model was used to generate short descriptions for each meme image, providing an additional modality that attempts to bridge the semantic gap.

A new version of the dataset was created by appending these generated image descriptions as an additional text column. The RoBERTa models were then fine-tuned using this augmented dataset, with all other training parameters held constant. The objective was to determine whether explicit image-level semantic hints could help the model better contextualize the meme content and thus improve toxic content detection.

MoonDream-Enhanced Data

To further enrich the textual modality, descriptions generated by the MoonDream model were integrated into the training data. These descriptions, derived from the visual content of the memes, aimed to provide additional semantic cues that the original caption text alone might not convey. This approach was motivated by understanding that meme comprehension often arises from the interplay between image and text, rather than from either modality in isolation.

As presented in Table 10, the inclusion of MoonDream-generated descriptions resulted in measurable performance gains across both RoBERTa base and large models. In particular, the F1 scores for the toxic class improved, indicating better model sensitivity to minority class samples. The base model achieved an F1 score of 0.5608 for the toxic class, while the large model closely followed with 0.5794. Accuracy also improved slightly for both models compared to the previous configuration.

These results validate the hypothesis that providing richer textual context, even when synthetically generated, can aid transformer-based models in toxic content classification tasks. It further underscores the value of multimodal alignment and semantic enhancement in the meme understanding task.

Table 10. RoBERTa-Finetuned with MoonDream Description, Learning Rate Optimization and 3× Oversampling Result.

Base Model	Precision	Recall	F1	Support
Non-toxic	0.9359	0.9380	0.9370	1775
Toxic	0.5652	0.5564	0.5608	257
Accuracy			0.8898	2032
macro avg	0.7506	0.7472	0.7489	2032
weighted avg	0.8890	0.8898	0.8894	2032
Large Model	Precision	Recall	F1	Support
Non-toxic	0.9379	0.9431	0.9404	1775
Toxic	0.5911	0.5681	0.5794	257
Accuracy			0.8957	2032
macro avg	0.7645	0.7556	0.7599	2032
weighted avg	0.8940	0.8957	0.8948	2032

T5 Paraphrasing Implementation

In the final text-based enhancement, paraphrases were generated using the T5 model and added to the original caption and MoonDream descriptions to increase textual equivalent, yet syntactically varied, textual forms. However, as shown in Table 11, this approach led to a noticeable degradation in performance, particularly in the toxic detection.

For the base model, the F1 score for the toxic class dropped to 0.3138, a sharp decline from 0.5608 achieved with MoonDream-only augmentation. Similarly, the large models' toxic F1 score decreased to 0.5126, despite maintaining relatively high accuracy in non-toxic prediction. These results suggest that paraphrasing introduced noise and semantic drift, diluting the toxic signals present in the original caption and descriptions.

Table 11. RoBERTa-Finetuned with MoonDream Description, Paraphrased, Learning Rate Optimization and 3× Oversampling Result.

Base Model	Precision	Recall	F1	Support
Non-toxic	0.8965	0.9662	0.9300	1775
Toxic	0.4958	0.2296	0.3138	257
Accuracy			0.8730	2032
macro avg	0.6961	0.5979	0.6219	2032
weighted avg	0.8458	0.8730	0.8521	2032
Large Model	Precision	Recall	F1	Support
Non-toxic	0.9295	0.929	0.9293	1775
Toxic	0.5116	0.5136	0.5126	257
Accuracy			0.8765	2032
macro avg	0.7206	0.7213	0.7209	2032
weighted avg	0.8767	0.8765	0.8766	2032

The training procedure for this configuration took approximately 1 h and 13 min per epoch, totalling 24 h and 20 min over 20 epochs. The increased computational effort, performance collapsed, indicating that excessive paraphrasing may have generalized away crucial linguistic features that differentiated toxic from non-toxic memes.

This outcome implies a saturation point for textual augmentations within the current dataset size and class imbalance. It also reinforces the need to move beyond purely text-based strategies, as further textual manipulation appears to offer diminishing returns for this classification task.

4.2. Vision Transformer Experiments

The Vision Transformer experiments focus on the visual modality extracted from both the Facebook and Innopolis Hateful Meme datasets. In these experiments, the model is provided solely with the image component of the meme. The classification task relies exclusively on the visual features to determine whether an image contains hateful (1) or non-hateful (0) content.

The primary objective is to fine-tune the Vision Transformer for toxic content detection and to evaluate how much the visual modality contributes to the overall classification. These understanding are subsequently applied to enhance the performance of the multimodal RoBERTViT model. The experimental procedure involves progressive data augmentation and the application of a custom loss function, aimed at improving classification performance, particularly for the underrepresented hateful class.

4.2.1. Vision Transformer on Facebook Hateful Memes

To assess the effectiveness of visual features in a controlled environment, a Vision Transformer model was fine-tuned using only the image modality from Facebook Hateful Meme dataset. Unlike the Innopolis Hateful Meme dataset, the Facebook Hateful Meme dataset is constructed with “benign confounder”, meaning the images are often harmless on their own and only become toxic when paired with specific text. This makes the vision-only task exceptionally challenging.

Training pipeline was implemented using PyTorch and the Hugging Face Transformer library. Image processing was conducted using ViTImageProcessing to standardize the input format. Each image was resized to 224×224 pixels for base model and 384×384 for the large model, in accordance with the model architecture requirements.

Training was performed on a system equipped with an NVIDIA RTX A6000 GPU. Both the base and large ViT models were trained using the Adam optimizer with a learning rate of 1×10^{-5} , and the cross-entropy loss function served as the training objective. Each model was trained over 20 epochs. During each epoch, the model was evaluated on the development set to monitor for overfitting.

Baseline Result

The initial experiments involved fine-tuning Vision Transformer models on the original Facebook Hateful Meme dataset without augmentation. The results are summarized in Table 12 for the base baseline model, and Table 13 the large baseline model.

Table 12. Finetuned ViT-Base on Facebook Hateful Meme Dataset.

Base Model	Precision	Recall	F1	Support
Non-toxic	0.5321	0.7961	0.6379	510
Toxic	0.5612	0.2714	0.3659	490
Accuracy			0.5390	1000
macro avg	0.5466	0.5338	0.5019	1000
weighted avg	0.5464	0.5390	0.5046	1000

Table 13. Finetuned ViT-Large on Facebook Hateful Meme Dataset.

Large Model	Precision	Recall	F1	Support
Non-toxic	0.5391	0.6353	0.5833	510
Toxic	0.5338	0.4347	0.4792	490
Accuracy			0.537	1000
macro avg	0.5365	0.535	0.5312	1000
weighted avg	0.5365	0.537	0.5323	1000

The result in Tables 12 and 13 highlight the difficulty of the task. The accuracy hover around 54%, which is only marginally better than random guessing in a balanced binary classification task. This confirms that for the Facebook Hateful Meme dataset, the visual modality alone carries very little toxic signal, as the dataset was specifically designed to require multimodal reasoning.

Data Augmentation

To attempt to extract more robust features, standard image augmentation techniques (flipping, rotation, color jittering) were applied. The result are presented in Tables 14 and 15.

Table 14. Result on Finetuned with Augmented Input on ViT-Base (Facebook Hateful Meme).

Base Model	Precision	Recall	F1	Support
Non-toxic	0.5114	0.702	0.5917	510
Toxic	0.4933	0.302	0.3747	490
Accuracy			0.506	1000
macro avg	0.5024	0.502	0.4832	1000
weighted avg	0.5026	0.506	0.4854	1000

Table 15. Result on Finetuned with Augmented Input on ViT-Large (Facebook Hateful Meme).

Large Model	Precision	Recall	F1	Support
Non-toxic	0.5329	0.7784	0.6327	510
Toxic	0.5569	0.2898	0.3812	490
Accuracy			0.539	1000
macro avg	0.5449	0.5341	0.5069	1000
weighted avg	0.5446	0.539	0.5095	1000

Result on Tables 14 and 15 shows that, unlike RoBERTa experiments, image augmentation provided negligible or even negative returns for this dataset. The toxic F1-score for base model improved slightly, but the large model saw a performance regression. This reinforces the conclusion that the limitation is not a lack of visual variety, but the absence of explicit toxic cues in the image themselves.

Loss Function Optimization

To encourage the model to focus on harder examples, the Cross-Entropy loss was replaced with Focal Loss. The result are shown in Tables 16 and 17.

Result on Tables 16 and 17 show that the application of Focal Loss gain the most significant improvement for the Large model, raising the Toxic F1-score to 0.5309. This suggest that while the overall signal is weak, the model can be forced to learn some subtle visual pattern (possibly context or specific imagery associated with hate groups) when penalized heavily for missing them.

Table 16. Result on Finetuned, Augmented Input + Custom Loss Function on ViT-Base (Facebook Hateful Meme).

Base Model	Precision	Recall	F1	Support
Non-toxic	0.5365	0.7196	0.6147	510
Toxic	0.5475	0.3531	0.4293	490
Accuracy			0.54	1000
macro avg	0.542	0.5363	0.522	1000
weighted avg	0.5419	0.54	0.5239	1000

Table 17. Result on Finetuned, Augmented Input + Custom Loss Function on ViT-Large (Facebook Hateful Meme).

Large Model	Precision	Recall	F1	Support
Non-toxic	0.5264	0.4686	0.4959	510
Toxic	0.5037	0.5412	0.5309	490
Accuracy			0.514	1000
macro avg	0.515	0.5149	0.5134	1000
weighted avg	0.5153	0.514	0.513	1000

4.2.2. Vision Transformer on Innopolis Hateful Memes

To assess the effectiveness of visual features in hateful meme detection, a Vision Transformer model was fine-tuned using only the image modality from the Innopolis Hateful Meme dataset. This dataset comprises meme images annotated with binary labels, hateful (1) or benign (0). The data were partitioned into training (80%), development (10%), and test (10%) subsets, while maintaining a balanced class distribution across splits.

The training pipeline was implemented using PyTorch and the Hugging Face Transformer library. Image preprocessing was conducted using the ViTImageProcessing, which standardize the input format for ViT models. Each image was resized to 224×224 pixels for the base model and 384×384 pixels for the large model, in accordance with the model architecture requirements.

Training was performed on a system equipped with an RTX A6000 GPU. Both the base and large ViT models were trained using the Adam optimizer with a learning rate of 1×10^{-5} , and the cross-entropy loss function served as the training objective. Each model was trained over 20 epochs. During each epoch, the model was evaluated on the development set to monitor for overfitting. Training and validation losses were recorded and plotted across epochs to visualize learning progression. These loss plots were saved for documentation and reproducibility.

Two model variants were trained: The ViT-Base model with patch size 16 and input resolution of 224×224 pixels, and the ViT-Large model with patch size 16 and input resolution of 384×384 pixels. The base model required approximately 40 min to train, while the large model took 7 h to train for 20 epochs. Both models were initially trained using default hyperparameters, and the best-performing checkpoints were saved for final evaluation on the held-out test set to assess generalization to unseen data.

Baseline Result

The initial experiments involved fine-tuning Vision Transformer models on the original Innopolis Hateful Meme dataset without applying data augmentation or a custom loss function. Both the base and large ViT models were trained using a learning rate of 1×10^{-5} . Training for the base model required approximately 40 min for 20 epochs, while the large model required approximately 7 h for 20 epochs. These runs serve as the baseline for evaluating the effectiveness of future enhancements.

Table 18 summarizes the performance metrics for the base model (ViT-base-patch16-224), while Table 19 presents the result for the large model (ViT-large-patch16-384).

As shown in Tables 18 and 19, both ViT models achieve high performance on non-toxic samples but struggle with correctly identifying toxic content. This pattern mirrors the observations from the RoBERTa experiments, highlighting the class imbalance issue and the visual subtlety of toxic memes that make classification more difficult.

These results serve as the baseline for further experimentation. It is hypothesized that incorporating synthetically generated toxic data through augmentation may enhance the model's ability to generalize and recognize hateful content more effectively.

Table 18. Finetuned ViT-Base on Innopolis Hateful Meme Dataset.

Base Model	Precision	Recall	F1	Support
Non-toxic	0.9090	0.9566	0.9322	1775
Toxic	0.5305	0.3385	0.4133	257
Accuracy			0.8784	2032
macro avg	0.7197	0.6476	0.6728	2032
weighted avg	0.8611	0.8784	0.8666	2032

Table 19. Finetuned ViT-Large on Innopolis Hateful Meme Dataset.

Large Model	Precision	Recall	F1	Support
Non-toxic	0.9182	0.9482	0.9329	1775
Toxic	0.5377	0.4163	0.4693	257
Accuracy			0.8809	2032
macro avg	0.7279	0.6823	0.7011	2032
weighted avg	0.8700	0.8809	0.8743	2032

Data Augmentation

To improve the generalization capability, several data augmentation techniques were applied, including random horizontal flipping, rotation, color jittering, and random resizing. These transformations were intended to introduce variability and mitigate overfitting to specific visual patterns within the original dataset.

As shown in Tables 20 and 21, the augmented training data led to a modest improvement in the detection of toxic samples. Specifically, both the base and large Vision Transformer models exhibited higher recall and F1-scores for the toxic class compared to the baseline result, indicating enhanced sensitivity to minority class features. Notably, these improvements were achieved without a significant drop in the performance for the non-toxic class or overall accuracy.

Table 20. Result on Finetuned with Augmented Input on ViT-Base (Innopolis Hateful Meme).

Base Model	Precision	Recall	F1	Support
Non-toxic	0.9127	0.9662	0.9387	1775
Toxic	0.6078	0.3619	0.4537	257
Accuracy			0.8898	2032
macro avg	0.7603	0.6640	0.6962	2032
weighted avg	0.8742	0.8898	0.8774	2032

Table 21. Result on Finetuned with Augmented Input on ViT-Large (Innopolis Hateful Meme).

Large Model	Precision	Recall	F1	Support
Non-toxic	0.9180	0.9645	0.9407	1775
Toxic	0.6228	0.4047	0.4906	257
Accuracy			0.8937	2032
macro avg	0.7704	0.6846	0.7156	2032
weighted avg	0.8806	0.8937	0.8837	2032

These results suggest that incorporating syntactic variance through data augmentation contributes positively to the model's ability to learn from a limited dataset. This strategy serves as a foundation step before introducing more sophisticated techniques such as a custom loss function or multimodal fusion.

Loss Function Optimization

To address the class imbalance in the Innopolis Hateful Meme dataset, the standard cross-entropy loss was replaced with focal loss. Focal loss is specifically designed to down-weight well-classified examples, as a result, emphasizing harder and misclassified samples during training. This makes it particularly effective in scenarios where one class, in this case (Toxic memes), is underrepresented.

The application of focal loss resulted in a modest but consistent improvement in the recall and F1-scores for the toxic class, as shown in Tables 22 and 23. The base Vision Transformer model achieved a toxic class F1-score of 0.4647, while the large model achieved 0.4587. These improvements, although limited, indicate that the model became slightly more sensitive to toxic features compared to the baseline with standard loss.

Table 22. Result on Finetuned, Augmented Input + Custom Loss Function on ViT-Base (Innopolis Hateful Meme).

Base Model	Precision	Recall	F1	Support
Non-toxic	0.9162	0.9549	0.9352	1775
Toxic	0.5604	0.3969	0.4647	257
Accuracy			0.8844	2032
macro avg	0.7383	0.6759	0.6999	2032
weighted avg	0.8712	0.8844	0.8757	2032

Table 23. Result on Finetuned, Augmented Input + Custom Loss Function on ViT-Large (Innopolis Hateful Meme).

Large Model	Precision	Recall	F1	Support
Non-toxic	0.9191	0.9346	0.9268	1775
Toxic	0.4890	0.4319	0.4587	257
Accuracy			0.8711	2032
macro avg	0.7041	0.6833	0.6927	2032
weighted avg	0.8647	0.8711	0.8676	2032

Overall accuracy remained relatively stable, confirming that gains in toxic class detection did not come at the expense of non-toxic performance. Nevertheless, the magnitude of improvement was marginal, suggesting that the primary bottleneck lies not in the loss function but in the limited diversity and representation of toxic examples within the dataset.

Consequently, it is concluded that the Vision Transformer models reach a practical level of optimization under the current data constraints. Further performance gains are likely to require enhanced dataset quality, either through the inclusion of more toxic samples or by employing synthetic generation techniques, rather than additional tuning of the hyperparameter.

4.2.3. Few-Shot and Cross-Dataset Evaluation

To further analyze the robustness and generalization capabilities of the Vision Transformer models, a few-shot learning evaluation was conducted.

It is important to note that the baseline configuration was explicitly chosen for this evaluation rather than the optimized “best model” that utilized augmentation and focal loss. This distinction is critical for two reasons.

1. **Intrinsic Generalization:** The aim is to assess the raw representational quality of the Vision Transformer architecture. The “best model” benefits from extensive dataset-specific optimization that may mask the models true few-shot learning capabilities.

2. Unbiased Transferability: Highly optimized model often suffer from over-specialization to their training distribution. By using the baseline, its ensure that the cross-dataset evaluation measures genuine domain adaptation capabilities rather than the artifacts of aggressive hyperparameter tuning.

The experiment assesses how well these baselines model perform when adapted with a limited number of example (zero, one, five, 50 and 100 shots) across both dataset. The result summarized in Table 24 reveal distinct behaviors between the dataset.

Table 24. Few-Shot Learning Evaluation of Vision Transformer Models (Baseline Configuration).

Model	Dataset	Shot Evaluation (Average F1-Score)				
		Zero-Shot	One-Shot	Five-Shot	50-Shot	100-Shot
ViT_FB_Base	Facebook	0.5040	0.5140	0.5020	0.5020	0.5077
	Innopolis	0.5472	0.5527	0.5765	0.6650	0.7313
ViT_FB_Large	Facebook	0.5360	0.5340	0.5387	0.5413	0.5373
	Innopolis	0.6407	0.6629	0.7115	0.7807	0.8136
ViT_Innopolis_Base	Facebook	0.5240	0.5300	0.5313	0.5333	0.5333
	Innopolis	0.7180	0.7379	0.7618	0.8023	0.8366
ViT_Innopolis_Large	Facebook	0.5100	0.5040	0.5220	0.5127	0.5180
	Innopolis	0.6604	0.7133	0.7061	0.7818	0.8168

Based on the Table 24 The models trained on Facebook Hateful Meme dataset (ViT_FB) exhibit poor performance when evaluated on the Facebook Hateful Meme dataset itself, with scores hovering around 0.50–0.53 across all shots. This reinforces the finding that the visual modality in the Facebook Hateful Meme dataset acts as a “benign confounder” and lacks explicit toxic signals. However, interestingly the ViT_FB_Large model transfers surprisingly well to the Innopolis dataset achieving a score of 0.8136 at 100 shot. This suggests that the large-scale visual feature learned from Facebook Hateful meme diverse imagery ca be effectively repurposed for the Innopolis Hateful Meme dataset, which contains more explicit visual cues.

On the contrary, models trained on Innopolis (ViT_Innopolis) perform strongly on their native dataset, with the Base model reaching 0.8366 at 100-shot. However, they fail to transfer this performance to the Facebook Hateful Meme dataset, remaining stagnant around 0.53. This confirms that while Innopolis contains learnable visual patterns, these patterns do not generalize to the harder, context-dependent nature of the Facebook Hateful Meme dataset.

4.3. RoBERTa + Vision Transformer (RoBERViT) Experiments

The RoBERViT model integrates both visual and textual modalities by combining the feature of RoBERTa and Vision Transformer (ViT). The aim of this experiment is to evaluate the potential of multimodal approach in enhancing toxic meme classification by leveraging the complementary strength of language and image representation.

RoBERViT adopts a late-fusion architecture, in which features extracted independently by RoBERTa from the text dataset and ViT from the image dataset are concatenated to form a joint representation. This combined vector is then passed through a series of fully connected layers to perform binary classification. Such a configuration enables the model to simultaneously consider contextual language cues and visual elements, which is essential for accurately identifying subtle or implicit toxic content often embedded in multimodal memes.

The integration of both modalities is expected to reduce the limitation observed in unimodal models, particularly in cases where toxicity is not apparent in either the text or image alone. The following section explains more about the experiments, result and comparative analysis of RoBERViT against its unimodal counterpart.

4.3.1. RoBERViT on Facebook Hateful Memes

For experiments conducted on the Facebook Hateful meme dataset, the previously described preprocessing techniques were retained, including image data augmentation, image captioning using the MoonDream1 model, and paraphrasing of captions with a T-5-base model. Each meme instance was paired with its corresponding image and enriched textual description, ensuring consistent alignment between visual and textual modalities during training.

Both of the base and large variants of RoBERViT were trained with a batch size of 16 for 20 epochs on Facebook Hateful Meme dataset. Training was conducted with several regularization and optimization strategies, including label smoothing, cosine annealing learning scheduling, and early stopping. The AdamW optimizer was used to update the model parameters.

Feature embeddings were extracted from pretrained RoBERTa and Vision Transformer (ViT) encoders. These embeddings were then concatenated and passed through two fully connected layers with ReLU activation and dropout regularization before reaching the final classification head. This architecture enables joint reasoning over both modalities, leveraging the semantic richness of text and the visual cues present in the image to improve the detection of toxic content.

RoBERViT Baseline

Before applying customized enhancement, a baseline for the RoBERViT model was established. Table 25 presents the performance metrics for both the base and large versions of RoBERViT using default configurations on the Facebook Hateful Meme dataset. In this baseline setup, the original unaltered Facebook Hateful Meme dataset was used, and no custom hyperparameters were applied. The architecture employed a simple late fusion approach in which feature embeddings from pretrained RoBERTa and Vision Transformer (ViT) models were concatenated and passed to a final classification head.

Table 25. RoBERViT-Finetuned Baseline Result (Facebook Hateful Meme).

Base Model	Precision	Recall	F1	Support
Non-toxic	0.5972	0.8431	0.6992	510
Toxic	0.7193	0.4082	0.5195	490
Accuracy			0.63	1000
macro avg	0.6558	0.6257	0.6093	1000
weighted avg	0.6546	0.63	0.6111	1000
Large Model	Precision	Recall	F1	Support
Non-toxic	0.6403	0.8902	0.7449	510
Toxic	0.8076	0.4796	0.6018	490
Accuracy			0.689	1000
macro avg	0.7239	0.6849	0.6733	1000
weighted avg	0.7223	0.689	0.6748	1000

When compared with the standalone RoBERTa model and the Vision Transformer model, the baseline RoBERViT model demonstrates distinct advantages. For the Large model, the F1-score for the toxic class reached 0.6018, which is a significant improvement over the unimodal text baseline (0.5094) and the unimodal vision baseline (0.4792). This confirms that multimodal fusion is effective for the Facebook Hateful Meme dataset, where toxicity often relies on the interaction between image and text.

Following this baseline, the next phase of experimentation integrates the enhancements previously shown to improve performance in standalone models. These include image

augmentation, the use of a custom loss function, learning rate scheduling, oversampling, and multimodal enrichment through image captioning using MoonDream1.

Tuned Hyperparameter

Following the best hyperparameter configuration identified for RoBERTa and Vision Transformer in earlier experiments, these optimal settings were applied to the proposed RoBERViT model. Table 26 presents the result for both the base and large version of RoBERViT on the Facebook Hateful Meme dataset.

Table 26. RoBERViT-Finetuned with MoonDream Description Result (Facebook Hateful Meme).

Base Model	Precision	Recall	F1	Support
Non-toxic	0.6477	0.8471	0.7341	510
Toxic	0.7658	0.5204	0.6197	490
Accuracy			0.687	1000
macro avg	0.7067	0.6837	0.687	1000
weighted avg	0.7055	0.687	0.678	1000
Non-toxic	0.6852	0.7255	0.7048	510
Toxic	0.6957	0.6531	0.6737	490
Accuracy			0.69	1000
macro avg	0.6904	0.6893	0.6892	1000
weighted avg	0.6903	0.69	0.6895	1000

The optimized RoBERViT-Large model achieved a toxic F1-score of 0.6737. This represents a substantial gain over the baseline configuration (0.6018), validating the effectiveness of integrating MoonDream caption and optimized hyperparameter in a multimodal setting.

Effect of T5 Paraphrasing on Multimodal Fusion

To assess whether the multimodal architecture could maintain performance when the specific phrasing of the meme text is altered, T5-based paraphrasing was applied to the Facebook Hateful Meme dataset. This experiment test if the visual context is strong enough to guide the classification when the textual cues are semantically preserved but lexically changed.

The result are presented in Table 27. Similar to the trends observed in the textual experiments, the inclusion of paraphrased input resulted in a performance decline.

Table 27. RoBERViT-Finetuned with MoonDream Description and Paraphrased Input Result (Facebook Hateful Meme).

Base Model	Precision	Recall	F1	Support
Non-toxic	0.6246	0.8175	0.7075	510
Toxic	0.7186	0.4898	0.5825	490
Accuracy			0.656	1000
macro avg	0.6716	0.6527	0.645	1000
weighted avg	0.6707	0.656	0.6463	1000
Large Model	Precision	Recall	F1	Support
Non-toxic	0.6383	0.8235	0.7192	510
Toxic	0.7368	0.5143	0.6058	490
Accuracy			0.672	1000
macro avg	0.6876	0.6689	0.6625	1000
weighted avg	0.6866	0.672	0.6636	1000

For the Large model, the toxic F1-score decreased from 0.6737 using only MoonDream descriptions to 0.6058. This drop suggests that even in a balanced dataset like Facebook Hateful Meme, which is designed to rely on multimodal interaction, the model still depends

significantly on specific surface-level textual patterns. When these patterns are disrupted by paraphrasing, the visual embedding from the Vision Transformer are not sufficient to fully recover the lost semantic information.

4.3.2. RoBERTViT on Innopolis Hateful Meme

For experiments conducted on the Innopolis Hateful Meme dataset, the previously described preprocessing techniques were retained, including image data augmentation, image captioning using the MoonDream1 model, and paraphrasing of captions with a T5-based model. Each meme instance was paired with its corresponding image and enriched textual description, ensuring consistent alignment between visual and textual modalities during training.

Both the base and large RoBERTViT configurations underwent training on Innopolis Hateful Meme dataset for 20 epochs using a batch size of 16. Parameter updates were managed by the AdamW optimizer, supported by a suite of optimization and regularization techniques that included a cosine annealing learning rate scheduler, label smoothing, and early stopping.

The architecture extracts feature embeddings utilizing pretrained Vision Transformer (ViT) and RoBERTa encoders. These distinct modalities are subsequently concatenated and processed through two fully connected layers—incorporating dropout regularization and ReLU activation—before feeding into the final classification head. By bridging the visual signals from the image with the semantic depth of the text, this joint reasoning framework enhances the model’s overall capacity to detect toxic content.

RoBERTViT Baseline

Prior to applying customized enhancement, a baseline for the RoBERTViT model was established. Table 28 presents the performance metrics for both the base and large versions of RoBERTViT using default configurations. In this baseline setup, the original (unaltered) Innopolis Hateful Meme dataset was used, and no custom hyperparameters were applied. The architecture employed a simple late fusion approach in which feature embeddings from pretrained RoBERTa and Vision Transformer (ViT) models were concatenated and passed to a final classification head.

Table 28. RoBERTViT-Finetuned Result.

Base Model	Precision	Recall	F1	Support
Non-toxic	0.9258	0.9707	0.9577	1775
Toxic	0.6959	0.4630	0.5561	257
Accuracy			0.9065	2032
macro avg	0.8109	0.7169	0.7519	2032
weighted avg	0.8968	0.9065	0.8982	2032
Large Model	Precision	Recall	F1	Support
Non-toxic	0.9335	0.9735	0.9531	1775
Toxic	0.7403	0.5214	0.6119	257
Accuracy			0.9163	2032
macro avg	0.8369	0.7475	0.7825	2032
weighted avg	0.9091	0.9163	0.9100	2032

When compared with standalone RoBERTa model (Table 7) and the Vision Transformer model (Table 18 for base and Table 19 for large model), the baseline RoBERTViT model already demonstrates improved performance in identifying toxicity, particularly in cases where harmful intent is implied through the interplay between visual and text elements.

Despite the absence of any customization, the multimodal fusion provides a noticeable advantage in terms of both F1-score and accuracy.

Following this baseline, the next phase of experimentation integrates the enhancements previously shown to improve performance in standalone models. These include image augmentation, the use of a custom loss function, learning rate scheduling, oversampling, and multimodal enrichment through image captioning using MoonDream1. The hypothesis is that incorporating these strategies into RoBERTViT will further improve performance gains by enhancing its ability to reason jointly over both modalities.

Tuned Hyperparameter

Following the best hyperparameter configuration identified for RoBERTa and Vision Transformer in earlier experiments, these optimal settings were applied to the proposed RoBERTViT model. The assumption was that hyperparameters optimized for unimodal RoBERTa and Vision Transformer models would also enhance the performance of the multimodal RoBERTViT architecture.

Table 29 presents the results for both the base and large versions of RoBERTViT. For the base model, the toxic F1-score reached 0.5708, representing an improvement over the RoBERTViT original configuration, as well as the RoBERTa best model shown in 10 and the base Vision Transformer model in Table 22.

Table 29. RoBERTViT-Finetuned with MoonDream Description Result.

Base Model	Precision	Recall	F1	Support
Non-toxic	0.9337	0.9515	0.9425	1775
Toxic	0.6143	0.5331	0.5708	257
Accuracy			0.8986	2032
macro avg	0.7740	0.7423	0.7567	2032
weighted avg	0.8933	0.8986	0.8955	2032
Large Model	Precision	Recall	F1	Support
Non-toxic	0.9418	0.9656	0.9535	1775
Toxic	0.7123	0.5875	0.6439	257
Accuracy			0.9178	2032
macro avg	0.8270	0.7766	0.7987	2032
weighted avg	0.9127	0.9178	0.9144	2032

The large model achieves a toxic F1-score of 0.6439, representing an improvement over RoBERTViT original configuration, as well as the RoBERTa best model shows in Table 10 and the large Vision Transformer model in Table 21.

Effect of T5 Paraphrasing on Multimodal Fusion

To evaluate the robustness of the multimodal model against linguistic variations, the T5-based paraphrasing strategy was applied to the best-performing RoBERTViT configuration. The hypothesis was that the visual feature from the Vision Transformer might compensate for the semantic drift introduced by text paraphrasing.

The Table 30 summarizes the results. Contrary to the hypothesis, the inclusion of paraphrased text caused a performance regression across both model sizes.

For the Large model, the toxic F1-score dropped from 0.6439 (MoonDream only) to 0.6027 (Paraphrased). This decline indicates that the multimodal model remains heavily reliant on specific textual patterns found in the original captions. When these patterns are altered by paraphrasing, the visual context provided by the Vision Transformer is insufficient to maintain classification accuracy.

Table 30. RoBERTViT-Finetuned with MoonDream Description and Paraphrased Input Result (Innopolis Hateful Meme).

Base Model	Precision	Recall	F1	Support
Non-toxic	0.9473	0.8907	0.9181	1775
Toxic	0.4656	0.6576	0.5452	257
Accuracy			0.8912	2032
macro avg	0.7064	0.7741	0.7316	2032
weighted avg	0.8863	0.8612	0.8709	2032
Large Model	Precision	Recall	F1	Support
Non-toxic	0.9337	0.9685	0.9508	1775
Toxic	0.7068	0.5253	0.6027	257
Accuracy			0.9124	2032
macro avg	0.8203	0.7469	0.7767	2032
weighted avg	0.9050	0.9124	0.9067	2032

5. Qualitative Analysis and Explainability

While quantitative metrics (F1-score) provide a high-level view of performance, they often obscure the underlying reasoning process. To address this, a qualitative analysis using Explainable (XAI) on specific case studies from the Facebook Hateful Meme dataset was conducted. This section compares the reasoning patterns of the text-specialist (RoBERTa), the vision-specialist (ViT), and the multimodal generalist (LLaVA).

5.1. Analysis on Facebook Hateful Meme Dataset

Three representative memes were selected from the Facebook Hateful Meme (FHM) dataset to illustrate distinct model behaviors: a correctly classified hateful meme, an incorrectly classified hateful meme, and a correctly classified non-hateful meme. Displayed in Figure 5 are the three specific examples FHM Meme 1, FHM Meme 2, and FHM Meme 3, which serve as the visual basis for the qualitative evaluation.

**Figure 5.** Selected Memes from Facebook Hateful Meme Dataset for Qualitative Analysis

5.1.1. Case 1: Explicit Toxicity (The Wheelchair Meme)

Figure 6 shows the analysis of a hateful meme targeting individuals with disabilities and race. This composite figure presents the original meme alongside the interpretability outputs for both the text and vision unimodal models, illustrating a distinct difference in how each modality processes the information:

- Figure 6 on LIME Text Analysis: The model correctly identified this as hateful (93.9% confidence). LIME visualization reveals the decision was driven by the keyword “black”, “white”, and “gang”. The model effectively acted as a keyword detector.
- Figure 6 on LIME Visual Analysis: The Vision Transformer also predicted “Hateful”, but LIME shows it did not look at the person in the wheelchair. Instead, it focused exclusively on the text overlay regions. The model learned to use the visual presence of text blocks as a proxy for hate.

- LLaVA (Generative): LLaVA correctly classified it as hateful. However, its visual analysis stated: “The visual element is a man in a wheelchair, which is an unrelated aspect”. This highlights a failure in multimodal integration, the model ignored the visual context that made the text ableist.

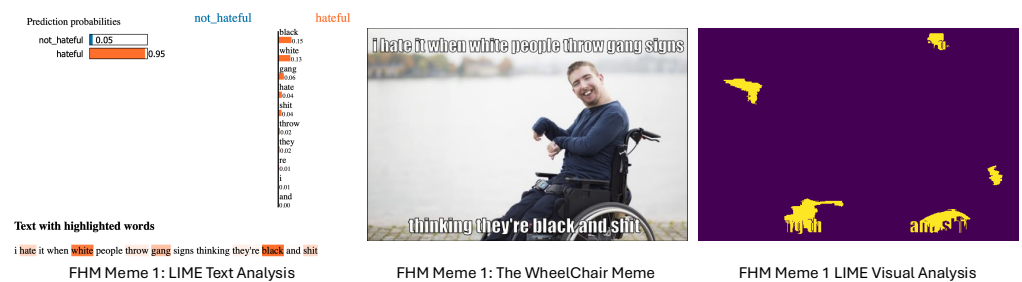


Figure 6. Analysis of Wheelchair Meme from Facebook Hateful Meme Dataset

5.1.2. Case 2: Contextual Toxicity (The 9/11 Meme)

Figure 7 shows the analysis of the Twin Tower burning with the caption “stupid neighbours woke me up”. This represents a “benign confounder” where the text is safe in isolation. The composite figure presents the original meme alongside the interpretability outputs for both the text and vision unimodal models, illustrating a distinct difference in how each modality processes the information:

- Figure 7 on LIME Text Analysis: The model failed, predicting “Non-Hateful”. LIME shows the word “stupid” strongly pushed the model toward the non-hateful class, masking the subtle toxicity of “woke me up”.
- Figure 7 on LIME Visual Analysis: The model also failed. The attention map was scattered across the sky and smoke, failing to identify the tragic nature of the event.
- LLaVA (Generative): LLaVA initially leaned toward non-hateful. Its explanation noted the “offensive language” but explicitly stated, “The image itself does not seem to be a factor”. This is a critical hallucination, the image is only factor that makes the meme hateful.



Figure 7. Analysis of 9/11 Meme from Facebook Hateful Meme Dataset

5.1.3. Case 3: Non-Hateful (The Orangutan Meme)

Figure 8 shows the analysis of a non-toxic meme discussing animal testing. The composite figure presents the original meme alongside the interpretability outputs for both the text and vision unimodal models, illustrating a distinct difference in how each modality processes the information:

- Figure 8 on LIME Text Analysis: Correctly predicted “Non-Hateful”. The model associated words like “intelligent” and “vegetable” with non-toxic content.
- Figure 8 on LIME Visual Analysis: Correctly predicted “Non-Hateful”, attending to the monkey’s arms and the text layout.
- LLaVA (Generative): LLaVA exhibit a reasoning-classification misalignment. While its quantitative confidence is 33.8% correctly placed it below the hate threshold, its

textual explanation hallucinated that “test drugs” were toxic terms. This confirms that the model’s text generation are hyper-sensitive, even when its internal probability distribution is uncertain.

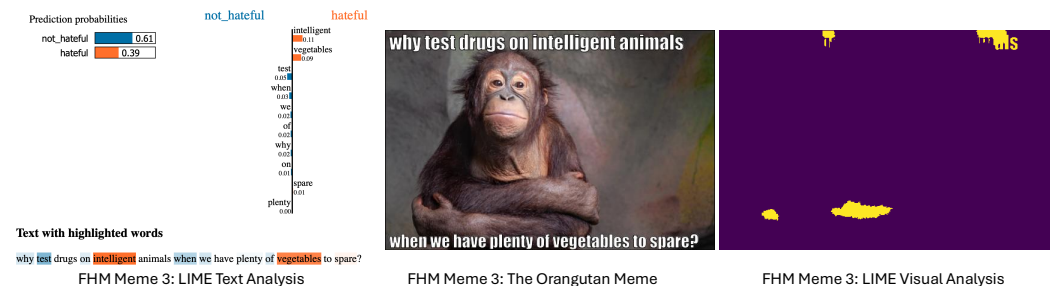


Figure 8. Analysis of Orangutan Meme from Facebook Hateful Meme Dataset

5.2. Analysis on Innopolis Hateful Meme Dataset

To understand how the models adapt to “in-the-wild” content, the qualitative analysis is extended to the Innopolis Hateful Meme (IHM) dataset. Unlike the Facebook Hateful Meme dataset, the Innopolis Hateful Meme dataset features a real-world distribution of memes where toxicity is often more explicit or culturally specific. Displayed in Figure 9 are the three specific examples IHM Meme 1, IHM Meme 2, and IHM Meme 3, which serve as the visual basis for the qualitative evaluation.



Figure 9. Selected Memes from Innopolis Hateful Meme Dataset for Qualitative Analysis

5.2.1. Case 1: Explicit Symbolism (The Hitler Meme)

Figure 10 shows the analysis of the Hitler Meme, which features an image of Adolf Hitler with the text “voice of Germany mein favorit”. The toxicity stems from the glorification of a historical figure associated with genocide and hate. The composite figure presents the original meme alongside the interpretability outputs for both the text and vision unimodal models, illustrating a distinct difference in how each modality processes the information:

- Figure 10 on LIME Text Analysis: The model correctly classified the meme as hateful with 67% confidence. The LIME analysis reveals that the decision was driven almost entirely by the word “Germany” and “mein”. This suggests the model act as a keyword detector, having learned a statistical association between these terms and hate speech within the specific context of this dataset, even though the words are neutral in isolation.
- Figure 10 on LIME Visual Analysis: The Vision Transformer also predicted the meme as hateful. However, the attention map does not focus on the defining facial feature of

the subjects. Instead, the highlighted regions align perfectly with the visual blocks of text (“voice”, “germany”, “mein”). This confirms the finding that the ViT relies on the presence and the visual shape of text overlays as proxy for hatefulness rather than recognizing the historical figure.

- LLaVA (Generative): LLaVA correctly classify the meme with confidence score of 67.6%. Unlike the Facebook Hateful Meme dataset case where it often ignored the image, here LLaVA explicitly integrated the visual modality, stating “The image of Adolf Hitler is a significant factor in the memes meaning” it reinforces the association between the quote and the historical figure. This indicates that for highly recognizable historical hate symbols, the multimodal model successfully grounds the text in visual reality.



Figure 10. Analysis of Hitler Meme from Innopolis Hateful Meme Dataset

5.2.2. Case 2: Hallucinated Toxicity (The World War II Meme)

Figure 11 shows the analysis of a depiction of a humorous meme referencing World War 2 with the caption “The allies start winning... ight imma head out”. The ground truth is non-hateful memes. The composite figure presents the original meme alongside the interpretability outputs for both the text and vision unimodal models, illustrating a distinct difference in how each modality processes the information:

- Figure 11 on LIME Text Analysis: The model correctly classified the meme as non-hateful with high confidence of 96.7%. The LIME analysis shows that the model did not find any strong toxic signals, words like “world” and “war” contributed slightly to the non-hateful prediction. This indicates that RoBERTa correctly identified the historical reference as non-hateful content rather than a threat.
- Figure 11 on ViT (LIME) Analysis: The Vision Transformer also correctly classified the meme as non-hateful. Consistent with previous observation, the attention map in highlights the random spots. Since the visual text features did not match the specific “hateful” font patterns or layouts the model learned during fine-tuning, it defaulted to the majority class of non-hateful, demonstrating that its correct prediction was likely due to the absence of a trigger rather than semantic understanding.
- LLaVA (Generative): In contrast to the specialists, the multimodal LLaVA model failed, incorrectly classifying the meme as Hateful. The explanation reveal a severe hallucination: the model claimed “the meme is hateful because it uses a derogatory term “crazy” to describe a situation. The word “crazy” does not exist in the memes text. Furthermore, it misinterpreted the slang “head out” as a “call to violence or aggression”. This highlights a critical failure mode in Large Multimodal Models where they may invent toxic features to justify a conservative safety decision”.

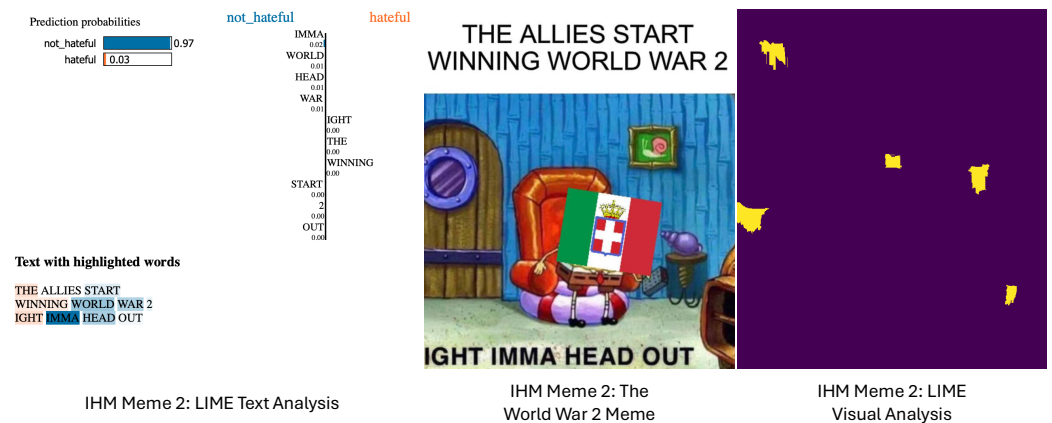


Figure 11. Analysis of World War II Meme from Innopolis Hateful Meme Dataset

5.2.3. Case 3: The Contextual Blind Spot (The “Wall” Meme)

Figure 12 shows the analysis of the “Wall” meme that contains the text referencing xenophobic tropes “when someone doesn’t reply to Spanish comments with ‘wall you have to go back’”. This relies on specific cultural knowledge of anti-immigrant rhetoric to be understood as hateful. The composite figure presents the original meme alongside the interpretability outputs for both the text and vision unimodal models, illustrating a distinct difference in how each modality processes the information:

- Figure 12 on LIME Text Analysis: The model completely failed to detect toxicity predicting “Non-Hateful” with 99% confidence. The LIME analysis reveals that the model was influenced by neutral words such as “spanish”, “reply”, and “with”. It treated the sentences as standard grammatical structure, failing to recognize that the combination of “spanish”, “wall”, and “go back” constitutes a specific hate.
- Figure 12 on LIME Visual Analysis: The Vision Transformer also classified the meme as non-hateful. The LIME visualization shows the model attention was concentrated on the visual text blocks, similar to the previous cases. However, since the font and layout did not match the visual “signature” of hate it had learned, the visual proxy failed to trigger a hateful prediction.
- LLaVA (Generative): LLaVA also failed to robustly identify the hate, classifying it with a low confidence of 32.4% (effectively predicting as non-hateful). The generative explanation rationalized the text as a “play on words” and a “humorous situation”, explicitly stating the meaning is “asking for a response to a comment in Spanish”. This demonstrates that even multimodal generalists can be misled by the structural format of a meme, interpreting hate speech as benign joke when the violence is implied rather than explicit.



Figure 12. Analysis of “Wall” Meme from Innopolis Hateful Meme Dataset

6. Comparative Analysis and Discussion

The experimental results provide several key insights into the challenges and opportunities of multimodal toxic meme detection. This section synthesizes the main discoveries, analyzes the dominance of the textual modality, compares results with prior work and outlines the limitations and ethical considerations of the study.

6.1. Summary of Findings

The comprehensive experiments reveal a clear hierarchy in model performance across both the Facebook Hateful Meme and Innopolis Hateful Meme dataset. Table 31 summarizes the best performance achieved by each architecture on the Facebook Hateful Meme dataset, and Table 32 summarized the best performance achieved by each architecture on the Innopolis Hateful Meme dataset.

Table 31. Summarized Results for Facebook Hateful Meme Dataset (Best Models).

	Precision	Recall	Accuracy	F1-Score
ViT-Base	0.5475	0.3531	0.5400	0.4293
ViT-Large	0.5037	0.5412	0.5140	0.5309
RoBERTa-Base	0.7584	0.6020	0.7110	0.6712
RoBERTa-Large	0.7586	0.6286	0.7200	0.6875
RoBERViT-Base	0.7658	0.5204	0.6870	0.6197
RoBERViT-Large	0.6957	0.6531	0.6900	0.6737

Table 32. Summarized Results for Innopolis Hateful Meme Dataset (Best Models).

	Precision	Recall	Accuracy	F1-Score
ViT-Base	0.5604	0.3969	0.8844	0.4647
ViT-Large	0.6228	0.4047	0.8937	0.4906
RoBERTa-Base	0.5652	0.5564	0.8898	0.5608
RoBERTa-Large	0.5911	0.5681	0.8957	0.5794
RoBERViT-Base	0.6143	0.5331	0.8986	0.5708
RoBERViT-Large	0.7123	0.5875	0.9178	0.6439

6.1.1. Facebook Hateful Meme Results

Table 31 summarizes the best performance achieved by each architecture on the Facebook Hateful Meme dataset.

On the Facebook Hateful meme dataset, the text-only RoBERTa-Large model achieved the highest F1-score of 0.6875. Slightly outperforming the multimodal RoBERViT-Large of 0.6737. The result supports the hypothesis that the visual modality in the Facebook Hateful Meme dataset often act as a “benign confounder”. The images themselves rarely contain explicit toxicity, and in some cases, adding the visual stream may introduce noise that marginally hampers the strong textual signal detected by RoBERTa.

The simple late-fusion architecture concatenates these conflicting visual signals, diluting the strong textual reasoning. This corroborated by the LIME analysis in Section 5.2, which revealed that on the Facebook Hateful Meme dataset, the ViT’s attention was often scattered across irrelevant background features (for example; the sky or the smoke) rather than focusing on the semantic subjects, thereby injecting noise into the fusion layer.

6.1.2. Innopolis Hateful Meme Results

In contrast, the results on the Innopolis Hateful Meme dataset, summarized in Table 32, demonstrate a clear advantage for the multimodal approach.

Table 32 shows the multimodal RoBERViT-Large model emerged as the clear top performer with a toxic-class F1-score of 0.6439, a significant improvement of over 6.4 points

compared to its best unimodal counterpart, RoBERTa-Large of 0.5794. This finding confirms that for the Innopolis Hateful Meme dataset, integrating visual features provided a consistent and substantial performance gain, suggesting that the visual modality in this dataset contains complementary cues that reinforce the textual prediction.

6.2. Lexical vs. Visual Dominance

Despite the gains from multimodal fusion, the result underscores the dominance of the textual modality for this specific task. A comparison between the RoBERTa results and the Vision Transformer results show that the Vision Transformer consistently struggles to identify toxic entries when used in isolation. This suggests that the image component contributes less to classification than initially assumed and that toxicity is primarily conveyed through language.

To further test this hypothesis, a series of classical machine learning experiments was conducted using a bags-of-words representation Tables 33 and 34.

Remarkably, simpler text-based models like Random Forest achieved competitive scores of 0.4891, outperforming the baseline Vision Transformer scores of 0.4647. This indicates that toxic memes can often be detected based on the presence of explicit toxic keywords.

This lexical dependence is further evidenced by the failure of the T-5 paraphrasing experiment. As shown in the RoBERTViT experiments, paraphrasing consistently harmed performance by altering the surface-level signals that the model depends on. This suggested that the transformer model did not achieve deep semantic understanding but instead learned to recognize specific toxic word patterns. Even with multimodal fusion, the visual signal was not strong enough to compensate for the loss of these textual triggers.

Table 33. Classic Machine Learning—Without Description.

Logistic Regression				
	Precision	recall	F1	support
Non-toxic	0.9102	0.9707	0.9395	1775
Toxic	0.6259	0.3385	0.4394	257
Accuracy			0.8907	2032
macro avg	0.768	0.6546	0.6894	2032
weighted avg	0.8742	0.8907	0.8762	2032
Naïve Bayes				
	Precision	Recall	F1	Support
Non-toxic	0.9132	0.9544	0.9333	1775
Toxic	0.5424	0.3735	0.4424	257
Accuracy			0.8809	2032
macro avg	0.7278	0.664	0.6879	2032
weighted avg	0.8663	0.8809	0.8712	2032
SVM (TF-IDF)				
	Precision	recall	F1	support
Non-toxic	0.9128	0.9735	0.9422	1775
Toxic	0.6619	0.358	0.4646	257
Accuracy			0.8957	2032
macro avg	0.7874	0.6657	0.7034	2032
weighted avg	0.8811	0.8957	0.8818	2032
Random Forest (TF-IDF)				
	Precision	recall	F1	support
Non-toxic	0.9131	0.9882	0.9491	1775
Toxic	0.8108	0.3502	0.4891	257
Accuracy			0.9075	2032
macro avg	0.8919	0.6692	0.7191	2032
weighted avg	0.9001	0.9075	0.891	2032

Table 34. Classic Machine Learning—With Description.

Logistic Regression				
	Precision	Recall	F1	Support
Non-toxic	0.913	0.9341	0.9234	1775
Toxic	0.4583	0.3852	0.4186	257
Accuracy			0.8647	2032
macro avg	0.6857	0.6596	0.671	2032
weighted avg	0.8555	0.8647	0.8596	2032
Naive Bayes				
	Precision	Recall	F1	Support
Non-toxic	0.9304	0.8811	0.9051	1775
Toxic	0.3989	0.5447	0.4605	257
Accuracy			0.8386	2032
macro avg	0.6646	0.7129	0.6828	2032
weighted avg	0.8632	0.8386	0.8489	2032
SVM (TF-IDF)				
	Precision	Recall	F1	Support
Non-toxic	0.9052	0.9735	0.9381	1775
Toxic	0.6179	0.2957	0.4	257
Accuracy			0.8878	2032
macro avg	0.7615	0.6346	0.6691	2032
weighted avg	0.8688	0.8878	0.8701	2032
Random Forest (TF-IDF)				
	Precision	Recall	F1	Support
Non-toxic	0.8977	0.9938	0.9433	1775
Toxic	0.8358	0.2179	0.3457	257
Accuracy			0.8957	2032
macro avg	0.8668	0.6059	0.6445	2032
weighted avg	0.8899	0.8957	0.8677	2032

6.3. The Paradox of Paraphrasing

A key finding of this study is the significant performance drop observed when T5-based paraphrasing was applied (Tables 27 and 30). While data augmentation typically improves robustness, here it reduced the toxic class F1-score for the Facebook Hateful Meme dataset from 0.6737 to 0.6058 for RoBERTa-Large model, and from 0.6197 to 0.5825 for RoBERTa-Base model. For Innopolis Hateful Meme is also the same toxic class F1-score is reduced from 0.6439 to 0.6027 for RoBERTa-Large model, and from 0.5708 to 0.5452 for RoBERTa-Base model.

This failure supports the hypothesis of Lexical Over-reliance. As revealed in the LIME analysis in Section 5.1, the text models operate primarily as sophisticated keyword detectors, relying on specific surface-level triggers (for example: “gang”, “stupid”, “vegetable”) rather than deep semantic intent. T5 paraphrasing preserves the meaning but alters the syntax and specific vocabulary. By removing the specific keywords the model “memorized” during training, the paraphrasing effectively blinded the model. This indicates that the models have not learned the underlying concept of hate, but rather a statistical dictionary of hateful term.

6.4. Comparison with State-of-the-Art Models

To contextualize our results, it is essential to compare them against previous benchmarks on the Innopolis Hateful Meme dataset. Table 35 provides a summary of our best-performing model alongside results from other studies.

To validate the effectiveness of the proposed RoBERTa framework, we compared its performance against several state-of-the-art benchmarks. This includes:

- LLaVA (Large Language and Vision Assistant): A state-of-the-art Large Multimodal Model evaluated in Zero-shot, 1-shot, 3-shot, and 5-shot settings to test generalization capabilities [24].

- Fine-tuned BERT Models: Specifically, the “bert-hateful-memes-expanded” model by Limjiayi [27] and the “autonlp-text-hateful-memes” model by am4nsolanki [28]. These represent strong text-only baselines fine-tuned on hate speech.
- Standard Visual Baseline: Fine-tuned ViT-Base and Res-Net-50 models (e.g., tommilyjones variants) were tested to establish pure visual classification performance.

Table 35 summarizes the best Toxic F1-scores achieved by these models compared to our RoBERViT-Large model on both dataset. For external models, the highest score achieved across the respective configuration was reported. (eg. with or without description, or optimal shot count for LLaVA).

Table 35. Comparison of Toxic Class F1-Scores with State-of-the-Art Models.

Model Type	Model Name	Facebook (Toxic F1)	Innopolis (Toxic F1)
Baselines	ViT-Base (tommilyjones)	0.4444	0.4508
	Limjiayi (BERT Expanded)	0.6722	0.5240
	am4nsolanki (AutoNLP)	0.5208	0.4636
Large Multimodal Model	LLaVA (Zero-shot)	0.6600	0.2365
	LLaVA (Few-shot Best *)	0.6247	0.2484
Proposed Method	RoBERViT-Large	0.6737	0.6439

* LLaVA Best: 3-shot for Facebook, 1-shot for Innopolis.

Table 35 shows striking insights into the nature of the datasets and the efficacy of the proposed architecture. On the Facebook Hateful Meme dataset, the RoBERViT-large model with F1-score of 0.6737 performs competitively with the strong BERT baseline by Limjiayi [27] with F1-score of 0.6772 and outperforms the Zero-shot LLaVA model with F1-Score of 0.6600. This confirms that on balances, standard benchmark, our late-fusion architecture is capable of matching the performance of large-scale pretrained system.

However, the result on the Innopolis Hateful Meme dataset highlight the specific strength of our approach.

- Failure of LLaVA: LLaVA performed poorly on Innopolis Hateful Meme dataset, with F1-scores hovering around 0.24 regardless of the number of shots (Zero to 5-shot). This suggests that Large Multimodal models, despite their size, struggle to grasp the specific, perhaps culturally nuanced or implicit toxicity present in the Innopolis Hateful Meme dataset without extensive fine-tuning.
- Superiority over Text Baselines: While Limjiayi [27] BERT model achieved a respectable F1-score of 0.5240, RoBERViT-Large significantly outperformed it with a F1-score of 0.6439. This gap over 12 points demonstrates that while text is dominant, the visual features integrated by RoBERViT provide crucial context that text-only models miss.

In summary, RoBERViT offer robust, specialized solution that outperforms general-purpose Large Multimodal Models and text-only baselines, particularly on complex, imbalanced dataset like Innopolis Hateful Meme dataset.

6.5. Study Limitations

The primary limitation of this study is the dataset itself. The Innopolis Hateful meme dataset is relatively small and highly imbalanced, with far more non-toxic samples than toxic ones. The scarcity of toxic examples restricts the generalization capability of large models like Transformers, increasing the risk of overfitting. While strategies like oversampling and focal loss were employed to mitigate this, the effectiveness is ultimately constrained by the available data.

Furthermore, as our analysis suggests, the dataset appears to be dominated by text-based toxicity. This may not fully represent the spectrum of hateful memes found online, where meaning is often derived from a subtle, implicit interplay between image and text. Consequently, the dataset may not be the ideal benchmark for showcasing the full potential of sophisticated multimodal fusion model.

6.6. Error Analysis and Models Failure

Across both datasets, three consistent error patterns emerged:

1. **The Sarcasm Gap:** As seen in the “9/11 meme” shown in Figure 7 case study, models consistently failed when the hate derived from ironic contrast. The text “stupid neighbors” is benign, and the image is tragic, but neither is explicit “hateful” in isolation. Late fusion fails to model this non-linear interaction.
2. **Benign Confounders:** On the Facebook dataset, the presence of high-quality, positive stock imagery (e.g., smiling faces) consistently lowered the confidence of hateful prediction acting as a visual “override” for the textual signal.
3. **Implicit Hate:** The models struggled with memes requiring external culture knowledge (e.g., the wall meme). Without an external knowledge base, the model treated the xenophobic as a literal statement.

6.7. Future Work

To address the limitation of small, culturally homogenous datasets, future work should focus on few criteria:

1. **Multilingual and Cross-Cultural Dataset:** Evaluating adaptability of hate to hate speech in diverse cultural context.
2. **Reasoning-Aware Architectures:** Moving beyond late fusion to Multimodal Chain-of-Thought (CoT) prompting, forcing models to explicitly articulate the relationship between image and text before classification.

7. Conclusions

This study presented a comprehensive comparative analysis of toxic meme detection using three distinct modeling architectures: a unimodal text specialist (RoBERTa), a unimodal vision specialist (Vision Transformer), and a proposed multimodal fusion framework, RoBERViT. To ensure a robust evaluation, the models were tested on two benchmarks with distinct characteristics: the imbalanced Innopolis Hateful Meme dataset and the balanced, confounder-driven Facebook Hateful Meme dataset. Additionally, a qualitative analysis using Explainable AI (XAI) and the LLaVA model was conducted to probe the underlying reasoning processes.

The empirical results reveal a complex relationship between modalities. On the Innopolis Hateful Meme dataset, the RoBERViT-Large model consistently outperformed its unimodal counterparts, achieving a significant performance gain with a toxic class F1-score of 0.6439, compared to the text-only baseline of 0.5794. This indicates that despite the dominance of text, the visual modality in Innopolis Hateful Meme contains complementary cues that reinforce classification. Conversely, results on the Facebook Hateful Meme dataset demonstrated that text-only models performed competitively with multimodal frameworks, supporting the hypothesis that “benign confounder” limits the effectiveness of straightforward visual features.

The qualitative analysis provided critical insights into why these models succeed or fail. RoBERTa functioned primarily as a sophisticated keyword detector, robust against explicit hate but brittle against sarcasm. The Vision Transformer largely failed to interpret semantic visual content, instead learning to use the visual presence of text overlays as

a statistical shortcut. Most notably, the LLaVA model, despite its generalist capabilities, exhibits a severe hallucination and a text-dominant bias, often misclassifying non-hateful memes based on imagined toxic features.

A key limitation observed was the detrimental effect of T5-based paraphrasing, which consistently degraded performance by diluting surface-level keywords. This confirms that current architectures suffer from lexical over-reliance rather than achieving deep semantic understanding.

Author Contributions: Conceptualization, N.S.A.B.N.A., M.P. and A.N.C.; methodology, N.S.A.B.N.A., M.P. and A.N.C.; software, N.S.A.B.N.A. and M.P.; validation, N.S.A.B.N.A., M.P., F.M. and A.N.C.; formal analysis, N.S.A.B.N.A. and M.P.; investigation, N.S.A.B.N.A. and M.P.; resources, M.P.; data curation, N.S.A.B.N.A.; writing—original draft preparation, N.S.A.B.N.A.; writing—review and editing, N.S.A.B.N.A., M.P., F.M. and A.N.C.; visualization, N.S.A.B.N.A.; supervision, M.P. and F.M. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not Applicable.

Informed Consent Statement: Not Applicable.

Data Availability Statement: Publicly available datasets were analyzed in this study. The Facebook Hateful Memes dataset can be found at [6], and the Innopolis Hateful Memes dataset can be found at [25].

Acknowledgments: The authors would like to thank the members of the Text Information Processing Laboratory at Kitami Institute of Technology for their valuable feedback and support. During the preparation of this manuscript, the authors used Gemini 3.1 Pro (Advanced, Google) for the purposes of proofreading, grammatical correction, and assisting with LaTeX code generation. The authors have reviewed and edited the output and take full responsibility for the content of this publication.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Devarapalli, L.T. Multimodal Hateful Meme Classification Using Vision Transformer and BERT. Master's Thesis, California State University, Sacramento, CA, USA, 2024.
2. Fortuna, P.; Nunes, S. A survey on automatic detection of hate speech in text. *ACM Comput. Surv. (CSUR)* **2018**, *51*, 1–30.
3. Schmidt, A.; Wiegand, M. A survey on hate speech detection using natural language processing. In Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media, Valencia, Spain, 3 April 2017; pp. 1–10.
4. Waseem, Z.; Hovy, D. Hateful symbols or hateful people? predictive features for hate speech detection on twitter. In Proceedings of the NAACL Student Research Workshop, San Diego, CA, USA, 13–15 June 2016; pp. 88–93.
5. Sharma, S.; Siddiqui, M.K.; Akhtar, M.S.; Chakraborty, T. Domain-aware self-supervised pre-training for label-efficient meme analysis. *arXiv* **2022**, arXiv:2209.14667.
6. Kiela, D.; Firooz, H.; Mohan, A.; Goswami, V.; Singh, A.; Ringshia, P.; Testuggine, D. The hateful memes challenge: Detecting hate speech in multimodal memes. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 2611–2624.
7. Ramos, G.; Batista, F.; Ribeiro, R.; Fialho, P.; Moro, S.; Fonseca, A.; Guerra, R.; Carvalho, P.; Marques, C.; Silva, C. A comprehensive review on automatic hate speech detection in the age of the transformer. *Soc. Netw. Anal. Min.* **2024**, *14*, 204.
8. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. *Adv. Neural Inf. Process. Syst.* **2017**, *30*, 1–11.
9. Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; Levy, O.; Lewis, M.; Zettlemoyer, L.; Stoyanov, V. Roberta: A robustly optimized bert pretraining approach. *arXiv* **2019**, arXiv:1907.11692.
10. Mozafari, M.; Farahbakhsh, R.; Crespi, N. A BERT-based transfer learning approach for hate speech detection in online social media. In *Proceedings of the Complex Networks and Their Applications VIII: Volume 1 Proceedings of the Eighth International Conference on Complex Networks and Their Applications COMPLEX NETWORKS 2019 8*; Springer: Berlin/Heidelberg, Germany, 2020; pp. 928–940.
11. Vidgen, B.; Derczynski, L. Directions in abusive language training data, a systematic review: Garbage in, garbage out. *PLoS ONE* **2020**, *15*, e0243300.
12. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv* **2020**, arXiv:2010.11929.

13. Touvron, H.; Cord, M.; Douze, M.; Massa, F.; Sablayrolles, A.; Jégou, H. Training data-efficient image transformers & distillation through attention. In *Proceedings of the International Conference on Machine Learning*; PMLR: Cambridge MA, USA, 2021; pp. 10347–10357.
14. Radford, A.; Kim, J.W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. Learning transferable visual models from natural language supervision. In *Proceedings of the International Conference on Machine Learning*; PMLR: Cambridge MA, USA, 2021; pp. 8748–8763.
15. Nguyen, D.Q.; Vu, T.; Nguyen, A.T. BERTweet: A pre-trained language model for English Tweets. *arXiv* **2020**, arXiv:2005.10200.
16. Ras, G.; Xie, N.; Van Gerven, M.; Doran, D. Explainable deep learning: A field guide for the uninitiated. *J. Artif. Intell. Res.* **2022**, *73*, 329–396.
17. Li, L.H.; Yatskar, M.; Yin, D.; Hsieh, C.J.; Chang, K.W. Visualbert: A simple and performant baseline for vision and language. *arXiv* **2019**, arXiv:1908.03557.
18. Lu, J.; Batra, D.; Parikh, D.; Lee, S. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *Adv. Neural Inf. Process. Syst.* **2019**, *32*, 1–11.
19. Kiela, D.; Bhooshan, S.; Firooz, H.; Testuggine, D. Supervised Multimodal Bitransformers for Classifying Images and Text. *arXiv* **2019**, arXiv:1909.02950.
20. Brown, T.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J.D.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; et al. Language models are few-shot learners. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 1877–1901.
21. Touvron, H.; Lavril, T.; Izacard, G.; Martinet, X.; Lachaux, M.A.; Lacroix, T.; Rozière, B.; Goyal, N.; Hambro, E.; Azhar, F.; et al. Llama: Open and efficient foundation language models. *arXiv* **2023**, arXiv:2302.13971.
22. Chowdhery, A.; Narang, S.; Devlin, J.; Bosma, M.; Mishra, G.; Roberts, A.; Barham, P.; Chung, H.W.; Sutton, C.; Gehrmann, S.; et al. Palm: Scaling language modeling with pathways. *J. Mach. Learn. Res.* **2023**, *24*, 1–113.
23. Zhang, X.; Lu, Y.; Wang, W.; Yan, A.; Yan, J.; Qin, L.; Wang, H.; Yan, X.; Wang, W.Y.; Petzold, L.R. Gpt-4v (ision) as a generalist evaluator for vision-language tasks. *arXiv* **2023**, arXiv:2311.01361.
24. Liu, S.; Cheng, H.; Liu, H.; Zhang, H.; Li, F.; Ren, T.; Zou, X.; Yang, J.; Su, H.; Zhu, J.; et al. Llava-plus: Learning to use tools for creating multimodal agents. In *Proceedings of the European Conference on Computer Vision*; Springer: Berlin/Heidelberg, Germany, 2024; pp. 126–142.
25. Badour, J.; Brown, J.A. Hateful memes classification using machine learning. In *Proceedings of the 2021 IEEE Symposium Series on Computational Intelligence (SSCI), Orlando, FL, USA, 5–7 December 2021*; IEEE: Orlando, FL, USA, 2021; pp. 1–8.
26. Ribeiro, M.T.; Singh, S.; Guestrin, C. “Why should i trust you?” Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, San Francisco, CA, USA, 13–17 August 2016; pp. 1135–1144.
27. Anghelina, I.; Buță, G.; Enache, A. SuteAlbastre at SemEval-2024 Task 4: Predicting Propaganda Techniques in Multilingual Memes using Joint Text and Vision Transformers. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, Mexico City, Mexico, 20–21 June 2024; pp. 443–449.
28. Singh, S.; Haridasan, A.; Mooney, R. “Female astronaut: Because sandwiches won’t make themselves up there”: Towards multimodal misogyny detection in memes. In *Proceedings of the 7th Workshop on Online Abuse and Harms (WOAH)*, Toronto, ON, Canada, 13 July 2023; pp. 150–159.

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.