Analyzing the Semantic Roles in Bangladeshi Memes: A Multimodal Approach to Contextual Role Labeling and Explanation Generation

by

Syed Mahbubun Nabi
22101472
Al Jami Islam Anik
22101577
Ahsan Shariar Khan Mazlish
22101635
Mustafis Ahsan
22101486

A thesis submitted to the Department of Computer Science and Engineering
in partial fulfillment of the requirements for the degree of
B.Sc. in Computer Science and Engineering

Department of Computer Science and Engineering
Brac University
January 2026.

Declaration

It is hereby declared that

1. The thesis submitted is my/our own original work while completing degree at Brac University.
2. The thesis does not contain material previously published or written by a third party, except where this is appropriately cited through full and accurate referencing.
3. The thesis does not contain material which has been accepted, or submitted, for any other degree or diploma at a university or other institution.
4. We have acknowledged all main sources of help.

Student's Full Name & Signature:

Mustafis Ahsan
22101486

Syed Mahbubun Nabi
22101472

Al Jami Islam Anik
22101577

Ahsan Shariar Khan Mazlish
22101635

Approval

The thesis/project titled "Analyzing the Semantic Roles in Bangladeshi Memes: A Multimodal Approach to Contextual Role Labeling and Explanation Generation" submitted by

1. Syed Mahbubun Nabi (22101472)
2. Al Jami Islam Anik (22101577)
3. Ahsan Shariar Khan Mazlish (22101635)
4. Mustafis Ahsan (22101486)

of Fall, 2024 has been accepted as satisfactory in partial fulfillment of the requirement for the degree of B.Sc. in Computer Science in Fall, 2025.

Examining Committee:

Supervisor:
(Member)

Dr. Jannatun Noor Mukta

Associate Professor, CSE
Director, BSDS
United International University

Co-Supervisor:
(Member)

Ariyan Hossain

Department of Computer Science and Engineering

Thesis Coordinator:
(Member)

Dr. Md. Golam Rabiul Alam

Professor
Department of Computer Science and Engineering
Brac University

Head of Department:
(Chair)

Dr. Sadia Hamid Kazi

Chairperson and Associate Professor
Department of Computer Science and Engineering
Brac University

Abstract

Interpreting the semantic roles of entities in memes requires navigating complex interactions between visual images, embedded texts, and implicit cultural knowledge. This is more challenging for languages with limited resources such as Bengali. Most existing works on this topic have only analyzed general concepts such as hate speech detection and sentiment analysis. They have not considered the detailed narrative roles that are required to comprehend the meaning of the memes. Our goal is to fill this gap and create an entity-based framework for semantic role labeling and humor explanation. To accomplish this, we have developed a labeled dataset of 2,658 Bengali memes, each labeled with roles such as Hero, Villain, and Victim, along with local context notes that were used to resolve cultural misunderstandings during training. We propose and evaluate two complementary frameworks: a Triple-Stream Fusion model utilizing cross-attention, and a Vision-Language Model (VLM) pipeline leveraging Qwen2-VL for holistic reasoning. Experimental results demonstrate that the VLM approach significantly outperforms the fusion baseline, achieving an accuracy of 0.8068 and a Macro-F1 score of 0.7953. Furthermore, our fine-tuned humor explanation generator achieved a significantly improved LLM-as-a-Judge quality score of 7.33/10 (compared to a 6.03 baseline), confirming its ability to produce culturally grounded interpretations. Our study is a strong benchmark on the task of examining satire and social commentary in Bengali memes and provides a simple approach to jointly label roles and generate explanations by combining images, text, and context.

Keywords: Bangla Memes. Contextualization, Multimodal, Computer Vision, Context-Aware Model

Table of Contents

List of Figures

Chapter 1

Introduction

1.1 Background

Introduction to meme
The word 'meme' first appeared in the book by Richard Dawkin [8], signified as a unit of transmission of culture. Memes, referred to as genes for cultures by him, tend to spread and vary through variation, selection and retention. Memes have become renowned over years, have changed its shape and form from text to picture and have become part of web communications in full swing in order to convey complex messages and shed light into social affairs.

Bengali memes and their characteristics
The language of Bengali, with stupendous diversity and with a million native speakers worldwide, is under-resourced in terms of computational tools for processing language naturally in an effective manner in terms of working with Bengali memes. There is stupendous diversity in language, and therefore, it is not an easier task to work with Bengali memes.

Importance of contextual understanding in meme
Memes in general can become pretty creative and vary a lot when expressing happiness, sarcasm, frustration, and many such feelings [22]. Thus, it is easy not to understand at all, even the real meaning of the meme. Besides, internet memes often lack attribution, allowing creators to distribute content freely, sometimes with questionable intent. This anonymity enables political and propaganda campaigns to manipulate narratives [2]. But given the context of a meme, it becomes easier to understand the targeted roles, characters and the true contents of a meme.

Challenges in meme explanation and context generation
Memes often implicitly ply sarcasm, satire, humor, irony and much more complex emotions and generating accurate explanations for such content is challenging due to the need for abstract reasoning and multimodal contextualization, as role labels in memes and corresponding explanations are interdependent on each other [38]. Also, it is even more challenging in Bengali due to the lack of adequate computational tools and not large enough datasets to train models effectively.

1.2 Rational of the Study or Motivation

In the web content universe, analysis of sentiment and explanation generation has become an important area of research. However, analysis of memes introduces an added level of difficulty with their multimodal content, nuanced nature, social event association, and cultural references in most of them. Hence, one must take into consideration the semantic and contextual background of a meme in order to interpret them in a proper manner [15]. The task turns out even more challenging in a low-resource scenario such as in the case of memes with Bengali-Text and with inbuilt local context in them. Because, there is a lack of benchmark datasets and not enough development in the local field [7]. Also, most of the works in the field of Meme-Analysis focus on naive factors such as hate speech and sarcasm classification but none with any semantic roles and the context in consideration. Semantic analysis of memes considering the contexts of the content, therefore, is an under-explored field, particularly for a low-resource language such as Bengali.

The motivation of this study is to enhance the interpretability and semantic comprehension of multimodal information, specifically in low-resource language environments such as Bengali. Despite being the most widely used language, Bengali content lacks the deserving scholarly attention in the area of multimodal semantics [33]. This research is conducted with a view to filling the gap in academic work regarding Bangla memes, a rich source for social critique and cultural expression. In addition, social-media use of memes as an effective tool for shared feelings and opinions raises a motivation for this study. It is a tool for political messaging and social critique [11] and semantic analysis of a Bangla Meme is important for an understanding of larger social trends and public feelings represented through such memes, providing useful insights for researchers, policymakers, and content moderators.

Therefore, the core motivation for this study is to bridge the research gap in multimodal semantic analysis for low-resource languages like Bengali. By developing a contextual role labeling and explanation generating model, this work seeks to make memes more interpretative, enable more effective sentiment analysis, and enable offensive content to be detected in a more effective manner.

1.3 Problem Statement

Memes are widely accepted to be an important digital communication means, mixing humor, political commentary, and social critique. Because memes convey complex messages in a combination of textual and non-textual things, they seem to be the most powerful instrument for entertainment and propaganda. In Bangladesh, memes are bound to represent socio-political settings and cultural values; they provide insight into generational perspectives on issues. However, the interpretation of memes is always complex due to their multimodal characteristics and the ugly syllable twists that apply codes from different languages, such as mixing Bengali and English [34], [42].

These challenges in explaining Bangladeshi memes are further complicated by the shortcomings of traditional computational models. Generally, text and image are processed separately in Natural Language Processing (NLP) and Computer Vision (CV) models, which may lead to the loss of contextual meaning that is core to meme comprehension [34], [42]. For instance, the interdependence of the textual and visual components is basic to constructing a meme's meaning, and this interdependence creates difficulties quite different from classical multimodal tasks such as image captioning [27]. Additionally, memes frequently use figurative language, sarcasm, or intertextual allusion, making them even more complex to analyze [35]. In the attempts involving powerful machine learning models have had a big trouble in making sense of semantic role labeling, which plays an integral role in defining how different components interact with one another in a meme to generate meaning [38].

The fluidity of memes complicates the situation since just one picture can incite an unthinkable number of meanings, depending on the caption, as well as the larger cultural discourse [34], [42]. Detecting such fluidity is typically hard for the machine learning models, which does not usually adjust its meaning depending on the shifting textual associations. The further cost of constantly adopting widely accepted state-of-the-art models for meme understanding would be invariably high, given the extensive training on immense datasets, aided by high-performance computational resources required to achieve a more significant model [34], [42]. This also serves as a hindrance towards greater access to myriad researchers and organizations, by virtue of money constraints, particularly in the given resource-constrained context of Bangladesh.

Moreover, machine learning models trained on global datasets face ethical concerns about their probable misinterpretation of region-specific political and cultural contexts, leading to an occasional mistake in identifying satire for misinformation or confirming prejudicial stereotypes [34], [42]. The proposed research intends, thus, to develop an integrated framework from multimodal data for Bangladeshi memes. It is constructed to draw on advanced deep learning techniques from natural language processing and computer vision with a focus on semantic role labeling and explanation generation [38]. The proposed research looks at developing contexts and modifying machine learning model capacity in understanding Bangladeshi memes to enhance digital literacy, miscommunication detection, and a deeper understanding of the shifting landscape of online conversation in the region [34], [42].

1.4 Objective

In our study, we aim to analyze the semantic roles in Bengali memes by using a multimodal approach integrating both visual and textual contents. This approach is to identify and also classify entities in memes such as heroes, villains or victims using the contextual role labeling techniques and examine how the interaction between texts and visuals builds meaning. Additionally, our goal is to establish a model that can generate context aware explanations and accurately analyze the semantic roles for local contexts. The main research objectives are:

1. To design a multimodal model able to take both text and image as input content present in a meme.

2. To identify and categorize the semantic roles (hero, villain, victim) of individuals or groups within a meme.

3. To generate context-aware relevant explanations for memes to understand the true intent and meaning of the meme.

4. To develop a rich dataset of annotated Bengali memes with cultural and contextual metadata.

5. To assist in fact-checking and content moderation by detecting biased or potentially harmful meme content.

1.5 Methodology in Brief

This work follows a multimodal entity-centric approach to comprehend Bangla memes through modeling the interdependency among visual content, embedded text, and associated metadata. It is informed by structured models of meme knowledge and inspired specifically by the Internet Meme Knowledge Graph, and aims to model how social roles and humor are constructed in interactions between image, language, and culture.

Research Approach
The study follows a mixed-methods, multimodal design that integrates computer vision and natural language processing to infer semantic roles (*hero, villain, victim*) at the entity level, and natural language modeling for generating grounded humor explanations. In contrast to simply propagating a single global label for each meme, the frameworld explicitly represents every labeled entity within it, allowing for more fine-grained analysis of how different targets are framed across multiple frames of them. The methodology focuses on quantitative role prediction as well as qualitative interpretability in terms of natural language explanations.

Data Collection
A Bengali meme dataset is constructed through structured collection and crowdsourced annotation from public social media groups and pages covering social, political, and cultural topics in Bangladesh. The dataset contains raw meme images along with OCR-extracted text, contextual fields (e.g., captions and descriptions), entity-level role labels, and human-written humor explanations. This design supports both supervised role classification and evaluation of explanation generation in a low-resource, culturally grounded setting.

Data Analysis
Three interrelated stages of analysis,representing the multimodal nature of memes:

Image-based Analysis
Pretrained vision transformers are employed to obtain global and token-level visual features of every meme image. These representations contain visual framing,

emphasis and scene composition information that often indicates how an entity is depicted.

Text-based Analysis
The textural signals are divided into two operative flows. We represent noisy and mixed OCR text (Bengali and English) using a Bengali-tuned language model to preserve its semantic clues, indicating praise, blame or ridicule. Topical and descriptive forms, as basically English alike form, are multilingually semantic encoded to denote higher framing and background knowledge.

Multimodal Fusion and Reasoning
Visual and textual representations are combined using attention-based fusion and, in a complementary pipeline, through end-to-end reasoning with a vision-language model. This design enables both modular inspection of modality contributions and holistic joint interpretation of image and text, which is essential for resolving sarcasm, implied targets, and culturally grounded humor.

Explanation Generation
A dedicated generation pipeline produces short, grounded natural-language explanations that describe why a meme is humorous and how the interaction between visual cues, OCR text, and context leads to a particular role framing. This step offers interpretation and aids in qualitative validation of predictions of the role.

Evaluation
The accuracy and macro-averaged accuracy of the role classification performance are assessed at the entity level based on the accuracy, recall, and F1-score to control the imbalance of the classes. Explanation quality is assessed using an LLM-as-a-Judge [24] framework with a fixed rubric, providing a scalable proxy for semantic alignment and groundedness. Uncertainty is quantified using statistical tests and confidence intervals, and the significance of observed differences between the variants of the models are assessed by using the same tools. [25].

Adaptation and Challenges
The methodology directly deals with low-resource and culturally specific memes content issues such as the problem of class imbalance, subjectivity in annotation, an OCR noise issue, and fast-changing meme trends. They are addressed in stratified data splitting, imbalance-sensitive training techniques, entity-based modeling, structured output constraints, and explanation generation in an attempt to enhance transparency and human interpretability.

1.6 Thesis Organization

Chapter-2 gives a review of literature that has been done on multimodal meme analysis, semantic role labeling and generation of the explanation, particularly on low resource and culturally based environments. Chapter 3 describes the system requirements, social and ethical concerns, and operational constraints according to which the proposed framework is designed. Chapter-4 describes the constructed Bangla meme dataset and presents the proposed methodology, including the triple-stream cross-attention classifier for role prediction, the vision-language model pipeline for structured role inference, and the humor explanation generation pipeline. The chapter-5 documents the experimental design, criteria of evaluation, statistical anal-

ysis and the comparison outcomes and the strengths and limitations of each design option are discussed in relation to previous literature. Lastly, Chapter-6 provides the summary of the principal findings of the thesis, major contributions to the existing body of knowledge in multimodal understandings of Bangali memes, and the future research directions.

Chapter 2

Literature Review

2.1 Preliminaries

Over the past decades, research on memes has significantly evolved, starting from understanding their textual and cultural impact to developing models for analyzing their multimodal nature [2, 37]. Multimodality is at the core of meme analysis, integrating multiple modes of communication like text and images to extract meaning [41]. In low-resource languages like Bengali, effective meme analysis requires a deep understanding of underlying methodologies and theories, particularly regarding language structure and the use of informal, code-mixed text [22, 29].

The global landscape for meme analysis has evolved with the use of deep learning, computer vision, and NLP approaches, particularly for tasks such as sentiment analysis, hate speech detection, and semantic role labeling [12, 36]. Techniques like Convolutional Neural Networks (CNNs), Vision Transformers (ViTs), and multimodal fusion methodologies are widely applied [26, 28].

For instance, CNNs effectively extract visual features from memes, while ViTs contribute to contextual semantic relations between visual and textual information [37, 41]. NLP techniques such as Named Entity Recognition (NER) and sentiment analysis further enhance meme text analysis, supplementing visual processing models [22, 23]. Frameworks such as MEMEX, VECTOR, and LUMEN capture context-awareness, entity extraction, and role classification, which are critical to meme interpretation [35, 38]. However, meme analysis still struggles with nuances, sarcasm, cultural specificity, and real-time adaptability to emerging trends.

When meme research is conducted in a local context, additional challenges arise, such as language mixing, complex cultural nuances, and the lack of benchmark datasets [1, 32]. Recent work on Bangla memes primarily employs XLM-RoBERTa and DenseNet-161 to detect hate speech, showing notable progress in classification [22, 29]. However, challenges such as dataset bias, overfitting, and the need for explainable AI techniques continue to affect model performance [10, 19]. These challenges highlight the necessity of developing frameworks that are culturally sensitive, adaptable, and capable of analyzing memes in low-resource languages like Bengali [20].

In summary, meme analysis in the context of semantic role labeling and explanation generation relies heavily on multimodal technologies, machine learning models, and semantic role labeling techniques [38]. By integrating these tools and addressing specific local challenges, it is possible to develop models that accurately interpret

complex roles and meanings in memes, ensuring cultural relevance and linguistic adaptability in AI-driven meme analysis [22, 36].

2.2 Review of Existing Research

Research on memes has increased over the past decade, particularly in describing how and why they work and mean, but most specifically in describing them in general and in detail. Memes are complex forms of media that combine images and language, conveying sophisticated messages through humor, satire, and cultural signifiers in an accessible manner. In our review, we examined work on language use in humans and general theory regarding analysis of memes in order to lay a strong background for our work.

Semantic Role Labeling in Memes

The task of semantic role labeling in memes has gained attention, particularly in identifying roles such as heroes, villains, and victims.Fharook et al. [12] utilizes name entity recognition, sentiment analysis, and similarity scoring to classify entities in memes, aiming to filter out toxic posts. Although their model achieves an F1 score of 23.855, it struggles with complex role definitions, indirect speech, and generalizability to new scenarios. This highlights the complexity of role labeling in memes and the need for flexible models that can comprehend nuanced language and context. Another similar study by Sharma et al. [36], uses the HVVMemes dataset—a collection of 7,000 political and COVID-19-related memes—examines the effectiveness of role labeling in a competitive setting. The top-performing model achieves an F1 score of 58.67, but shows bias toward the "other" category, has difficulty in detecting sarcasm, and a lack of common sense information persist. These findings underscore the importance of cultural and narrative awareness in meme analysis, as well as the need for context-aware models. Building on this, Sharma et al. [35] introduces the VECTOR model, a multimodal approach to role identification, combining images, text, and general information . Using the HVVMemes dataset, VECTOR outperforms existing models with a Macro-F1 score of 0.581, demonstrating a 4% improvement over unimodal models and a 12% gain over multimodal models. This work reveals that providing general information aids in discovering less prevalent roles, yet present technology continues to have a problem with discovering small contextual cues and infrequent categories. In a role classification comparison study, Sharma et al. [38] explore generating explanations for visual semantic role labeling in memes. They seek to detect such roles as heroes, villains, and victims. There are 4,680 entities with labels in 3,000 in its ExHVV dataset, and LUMEN, a model designed to generate explanations for detected roles, is presented in the study. It seeks to integrate both visual and textual information in generating effective role labels and explanations and in acknowledging strong datasets and models that work with a variety of meme types.

Sentiment and Emotion Analysis in Memes

Meme sentiment analysis is most difficult with integration of in-texted, image, and text. One work by Kumar et al. [23] proposes a context-aware model with integration of SentiBank, a lexicon-machine model, and CNNs for modalities' sentiment

polarity analysis in. As it is strong in analysis of in-text and image, but not in Optical Character Recognition (OCR), multi-language, and non-text modalities, current techniques have a long distance to cover in processing meme complexity. Liu et al. [26] proposes an ensemble model with integration of both text and video information for emotion analysis is proposed. With networks with incorporation of BERT, GPT-2, ResNet, and VGG, high accuracy (ACC-2: 84.1%, F1: 86.7%) is achieved in CMU-MOSI and (ACC-2: 84%, F1: 89.1%) in CMU-MOSEI datasets. It is strong in integration of a variety of modalities for emotion comprehension, but its application in memes is yet to be witnessed. In another work by Braytee et al. [3], social media posts in a 2023 Israel-Hamas war have been processed for hate and love expression with a Dual-Pipeline Attentional model. In its work, important trends such as posts expressing love with high probability for use of religious language and peace-themed emojis, and hate posts with high emoji use for expression of sad feelings and warnings, have been uncovered. In its contribution, it is beneficial in multimodal approaches in processing nuanced expression in political environments, but cultural and contextual diversity is a challenge in processing it. The role of political and affective environments with regards to memes is a field of study in its own right. For social posts regarding 2023 Israel-Hamas war,Ng et al. [31] explores a Dual-Pipeline Attentional model for emotion analysis is constructed in work in the Love-Hate Dataset. With a 73.18 accuracy level and an F1 score of 69.76 when trained and evaluated with MVSA-Multiple, it outperforms state-of-the-art techniques. What is focused in such work is analysis of emotive content with multimodal models, but not yet addressed is cultural and contextual nuances of memes.

Meme Detection and Template Identification

Beskow et al [2] introduces the Meme-Hunter model deals with political meme creation with a distinction between information that is a meme and information that is not a meme. With analysis of an image, Optical Character Recognition (OCR), and face detection, Meme-Hunter observes for memetics during and following 2018 Midterm Elections in America and 2018 National Elections in Sweden. That work is useful for following trends in an interval in a period of time with algorithms for learning in a graphical form, but not in real-time adaptability with new forms of emerging memes. Murg´as et al. [28] conduct a similar study with machine algorithms, including Convolutional Neural Networks (CNNs), distance-based classification, and clustering, for meme template analysis. Those algorithms can specifically respond to why and when a meme trends and impacts society, specifically in politics and advertisement. Nevertheless, perceptual hashing and CNNs lack in most cases. In that work, a need for flexible frameworks that can adapt with new trends in memes but preserve meaning is stressed.

Detection of Harmful or Critical Content in Memes

The identification of critical, dangerous, information in memes, including anti-vaccine disinformation and sexist information, is a new field of investigation. One such sork by Naseem et al. [30] introduces VaxMeme, an anti-vaccine tweet search

algorithm via both text and images. VaxMeme employs techniques such as attentive learning and ResNet-50 for image processing, with an 84.2% F1-score, a 2.55% improvement over state-of-the-art techniques. This work establishes the efficacy of multimodal techniques in critical information identification, but with a caveat about ease in generalizability towards new trends in social networks.Pasha and Wajeed [32] with similar objectives involves identifying sexist information in memes via state-of-the-art techniques in machine learning. Authors utilize the CLIP (Contrastive Language-Image Pre-training) model for developing Spanish and English datasets' multimodal representations. Authors utilize Random Forest, SVM, and Logistic Regression for testing and report best performance with Random Forest. Authors introduce work with high potential for employing high-end multimodal models for identifying dangerous information, but with a caveat about a necessity for multilinguality and diversity in datasets for generalizability improvement.

Contextualization and Explanation Generation for Memes

A large portion of meme studies is the Internet Meme Knowledge Graph (IMKG). Joshi et al. [21] proposes a mechanism for defining memes between websites . It identifies a similar background, picture, and cultural meaning with Vision Transformers. It performs exceedingly well (around 90%) with Reddit and Discord memes, but not with emerging trends not yet in IMKG. That issue reveals that we require alternative approaches to merge context and picture cues to comprehend a meme better. Building on this need for improved explanatory mechanisms, researchers have also explored methods for generating meme explanations by integrating visual, textual, and contextual information. The task of generating explanations for memes has also been explored, with a focus on integrating visual, textual, and contextual information. Sharma et al. [37] explores MEMEX, which proposes a way of obtaining explanations for memes by analyzing similar documents. This study constructs the Meme Context Corpus (MCC) of meme-sentence pairs explaining their context. The proposed MIME (Multimodal Meme Explainer) model combines visual, textual, and common sense knowledge through augmented encoders, Transformers, and LSTMs. This method illustrates the power of multimodal fusion to create relevant explanations, yet at the same time indicates the difficulty in processing vague or culturally relevant memes. In addition to explanation generation, ensuring fairness and transparency in VL models' explanations generated in meme analysis is also another key feature of meme analysis. Fairness and transparency of VL models' explanations generated in meme analysis have also been studied. Zhong et al. [41] investigates whether models such as LLaVA and MiniGPT-4 are able to provide transparent, as well as fair, explanations of memes. Through automated and manual inspection, the study identifies bias in the output, i.e., towards some visual signals or feelings, and a bias towards some roles such as villains, heroes, or victims. The paper highlights the necessity to remove biases in VL models to enable even and accurate meme analysis, though it also acknowledges the difficulty in doing so owing to the subjectivity entailed in interpreting memes.

Analyzing memes in a language such as Bangla, and in locally contextual settings, carries with it certain sets of requirements, such as accessing deeper meaning, cultural nuance, and expert domain knowledge. As such, these represent new avenues

for investigation, in terms of low-resource language such as Bengali. In follow-on work, relevant insights into current work in terms of Bengali memes and multimodal analysis of memes can then be derived, offering useful insights for current work.

Hate Speech, Offensive Content, and Cyberbullying Detection in Bengali Memes

One significant area of research is hate speech and offensive language detecting in Bengali memes. Karim et al. [22] constructs a model with image and textual information for hate speech in Bengali, a contribution to the field. This work constructs the largest publicly accessible hate speech dataset for Bengali and utilizes a deep architecture model with XLM-RoBERTa and DenseNet-161, with an accuracy of 0.83 for an F1-score. However, the study struggles with issues like overfitting with a relatively small dataset and interpretability for deep-rooted processes in complex meme structures. Nahin et al. [29] study constructs methodologies for hate speech in Bengali with deep approaches such as Xception, with 68.35% accuracy in a corpus of 3,000 memes. In its work, this study constructs multimodal analysis complications, namely in terms of interpretability of memes with cultural and contextual sensitivities. This work identifies useful insights in terms of utilizing multimodal architectures for semantic role labeling and interpretative analysis in Bangladeshi memes, with a focused role for combining contextual and feature-based methodologies. Identification of political violence and cyberbullying in Bangla texts forms yet a relevant direction for meme analysis. Hossain et al. [16] introduces DORA (Dual co-attention fRAmework), a hate and hate target model, a multimodal. It introduces a new dataset, 7,148 hate and target labeled Bengali memes, and utilizes a dual co-attention mechanism for simultaneous analysis of both textual and visual information and outperforms nine baseline methods with high generalizability for low-resource languages such as Hindi and Bengali. The work identifies strong effectiveness in combining multimodal cues for hate prediction but identifies weaknesses in processing implicitly present or culturally specific cues. Recommendations for future work include integration of larger contextual information and adversarial training. Ahmed et al. [1] introduces a variety of prediction algorithms for predicting cyberbullying in Bangla texts, with high accuracy values for binary (87.91%) and multiclass (79.29%) prediction. The study states problems such as lack of information, unbalanced classes, underlying sarcasm and satire detection difficulties. These issues match with similar concerns in multimodal satire analysis.This work is significant in underlining the imperative for robust contextual analysis techniques and complex model architectures and useful for the ongoing thesis work. In a study by Islam et al. [19], a weighted ensemble model for offensive Bengali meme detection is introduced, combining BERT for text classification with architectures for image classification. With use of the 3,662-meme-containing Multi-modal Offensive Memes Bengali Dataset (MoMBD), the model reaches a weighted F1-score of 71.84% for fine-grained classification. This work is significant in demonstrating the utility of multimodal integration for in-depth meme analysis and necessitates combining visual and textual information for semantic role labeling in Bangladeshi memes.

Sentiment Analysis and Explainability in Bengali Memes

The field of sentiment analysis in Bengali memes is significant, particularly since less-resourced languages suffer with it. A study by Jannat et al. [20], introduces a model that utilizes both text and images (VGG19 and BiLSTM) for sentiment analysis in Bengali memes reaches an F1-score of 0.68 with 1,671 samples of memes. Best performance happens when both sources of information are utilized together, not individually alone. According to this work, utilizing both texts and pictures for analysis is important for combining features, classification, and explaining meanings. Elahi et al. [10] considers explainable multimodal sentiment analysis in Bengali memes with MemoSEN, unifying text processing with BanglishBERT and image processing with ResNet50. It reaches a weighted F1-score of 0.71, proving multimodal models can effectively interpret memes, particularly when role information signifying sentiment is extracted. Yet, predicting a meme's sentiment when it is not explicit (a "no-opinion" meme) proves challenging in this work with unbalanced limited datasets. With Explainable AI (XAI) techniques such as LIME, a model can become easier to explain, even with uncertain and no-opinion memes.

In conclusion, the worldwide scenario of multimodal analysis for a meme reveals a vast number of gaps, such as real-time adaptability, dealing with changing meme structures, and model interpretability improvement. All these are even relevant when taking into consideration cultural and language-related aspects of a meme in a specific country, in our case, Bangladesh. Locally, in a review of work, we reveal an opportunity in dealing with these gaps in a scenario of a meme in Bangladesh. By taking cues from worldwide studies and overcoming specific obstacles in a meme analysis in Bangladesh, our thesis seeks to make a contribution in developing a strong multimodal model for contextual role labeling and explanation creation for Bengali memes. In filling a critical gap in work, in addition to a contribution, our work will present a sensitive and culturally-aware view in dealing with complex semantic roles in a meme.

2.3 Summary of Key Findings

Authors	**Year**	**Method Used**	**Dataset Used**	**Results**
Hossain et al. [16]	2024	DORA (Dual Co-Attention Framework)—multimodal deep neural network	BHM (Bengali Hateful Memes)	DORA outperformed baselines with F1 scores: 0.718 (hate detection), 0.720 (target identification).

Authors	Year	Method Used	Dataset Used	Results
Joshi et al. [21]	2024	Vision Transformer (ViT) for visual similarity matching; grounding memes to IMKG; ETL pipeline for data collection	13,175 Reddit memes; 15,985 Discord memes	Mapped memes to IMKG with high precision (90%+). Identified popular templates; 17.75% of Reddit image posts and 16.18% of Discord image posts were memes.
Sharma et al. [37]	2024	MIME (Multimodal Meme Explainer), MAT, and MA-LSTM for cross-modal contextualization	MCC (Meme Context Corpus): 3,400 memes with related context documents, annotated with evidence sentences	MIME achieved F1 = 0.813, outperforming baselines; improved accuracy (5.34%), precision (4.26%), recall (2.31%), and exact match (8.00%) over the best baseline.
Sharma et al. [38]	2023	LUMEN uses models like ViT, DeBERTa, and T5	ExHVV (3,000 memes)	LUMEN achieved the best performance across 18 NLG metrics, with BLEU-4 = 0.313 and CIDEr = 1.380.
Sharma et al. [35]	2023	VECTOR (Visual-semantic role dEteCToR): multimodal framework integrating entity context, commonsense knowledge, and optimal-transport kernel embedding	HVMemes (US Politics and COVID-19 memes)	VECTOR achieved Macro-F1 = 0.581, outperforming baselines and shared-task submissions.

Authors	Year	Method Used	Dataset Used	Results
Beskow et al. [2]	2020	Meme-Hunter: multimodal deep learning with OCR, face detection, and graph learning	25,109 meme images; 25,100 non-meme images; Twitter data from 2018 US Midterm Elections and 2018 Swedish National Elections	Meme-Hunter achieved 8x higher recall than template-based methods; F1 = 0.961 (multimodal model).
Karim et al. [22]	2023	Bi-LSTM, Conv-LSTM, XLM-RoBERTa, ResNet-152, DenseNet-161, multimodal fusion	Bengali Hate Speech Dataset (4,500 labeled memes)	F1 scores: 0.78 (Conv-LSTM), 0.82 (XLM-RoBERTa), 0.79 (DenseNet-161), 0.83 (XLM-RoBERTa + DenseNet-161).
Jannat et al. [20]	2022	BiLSTM, VGG19, CNN, multimodal fusion	Bengali meme dataset (1,671 memes)	F1-score: 0.68 (BiLSTM + VGG19).
Elahi et al. [10]	2024	Multimodal approach using ResNet50 and BanglishBERT	MemoSEN dataset (4,372 Bengali memes with captions)	Weighted F1-score: 0.71 (multimodal), 0.70 (DenseNet161), 0.66 (BanglishBERT).

Table 2.1: Key Findings

Research on memes has increased worldwide to explore how words and images correlate together. However, there exist certain limitations that reveal significant areas in need of further study. One significant limitation in this field of study is that new trends in memes are challenging for models to detect with accuracy. Internet Meme Knowledge Graph (IMKG) and templated-based models cannot adapt to new trends in a timely manner and become less effective over time [21, 28]. In addition, these trained models do not work effectively when being applied in new environments, such as political and COVID-19 meme datasets, for which they have been trained performs poorly on memes of other contexts [12, 36]. Not being able to utilize rich datasets of various contexts and categories, particularly for languages with scarce resources such as Bengali, reduces the effectiveness of the trained model [20, 22]. Another issue is that it is challenging to extract hidden underlying meanings from memes, such as sarcasm and cultural cues, since these depend on common information which the meme datasets lack [23, 35]. In addition, computational constraints reduce multimodal model capabilities and real-time application [2, 26].

The primary endeavor of this study is to address these above-mentioned limitations

which present a severe challenge in this field. There is a scarcity for meme datasets with various geographical and cultural contexts, integrated with general and commonsense knowledge, advanced multi modal models that can excel in real time accurate explanation generation. There is an opportunity to study and significantly improve multimodal meme analysis, explanation and therefore, the thesis aims at accurate semantic role analysis of memes and generating context-aware explanations,overcoming the need of local contexts to understand region specific contents, and developing robust, culturally sensitive, and scalable frameworks for meme analysis

In the local context, Bengali meme studies have mostly been in the area of classification tasks such as hate speech detection and sentiment analysis with multimodal methodologies. But there are significant pitfalls to these studies that point towards major gaps in the research. Bengali meme studies have mostly been in the area of classification tasks such as hate speech detection and sentiment analysis but there are significant pitfalls to these studies. Limited availability of varied datasets and use of small, biased datasets limit model generalizability and robustness [20, 22]. Moreover, excessive focus on categorization prevents further semantic comprehension, whereas models fail to comprehend implied meaning such as sarcasm and cultural references because of the lack of contextual and common-sense knowledge incorporation [10, 29]. Computational limitations also hamper scalability along with real-time deployment [10, 20].

These limitations identify research gaps that perfectly align with the current thesis theme. Through its emphasis on semantic role analysis, contextual and commonsense knowledge, and explanation generation, the thesis closes the gap between international developments and local needs towards enabling strong, culturally suitable, and scalable meme interpretation systems. Worldwide, models cannot handle new formats, lack diverse datasets, and cannot detect subtle meanings such as sarcasm, while computational limits prevent scalability . In Bengali, overdependence on small, unbalanced datasets and overuse of classification are standing in the way of semantic understanding at a deeper level and real-time scalability. These gaps indicate the requirement of culturally representative datasets, increased contextual understanding, and explanation generation. The present thesis addresses these by first giving priority to semantic roles, drawing inspiration from multimodal theories, and designing well-formed explanations, conjoining global advances with local requirements for robust, culturally sensitive systems.

Chapter 3

Requirements, Impacts and Constraints

3.1 Final Specifications and Requirements

Our project delivers an end-to-end research prototype for Bangla meme understanding that treats a meme as a genuinely multimodal artifact rather than plain text. In practice, each input instance consists of the meme image along with the textual signals that users typically rely on to interpret it, such as OCR-extracted text and available contextual metadata. Given this input, the system performs entity-centric interpretation: for a specified target entity, it predicts the entity's semantic role in the meme narrative (*hero, villain, victim, other*) and produces a short explanation that summarizes the cues behind the decision. Because meme text is frequently noisy and code-mixed, the pipeline is designed to remain usable even when OCR quality is imperfect or when optional fields are absent. The final design also emphasizes reproducibility and fair evaluation: dataset splits and preprocessing steps are fixed and logged so that results can be replicated, and the model is evaluated with imbalance-aware metrics so that performance is not overstated by majority-class behavior. From a practical standpoint, the prototype is intended to run in standard research environments, where GPU resources are preferred for training and representation extraction but the overall pipeline remains accessible for small-scale testing and analysis on CPU.

3.1.1 Scope of the Prototype

The prototype targets role prediction and explanation generation under realistic Bangla meme conditions, including mixed Bangla-English text, informal spelling, and culturally contextual references. The intended usage is research and analysis; it is not positioned as a fully automated decision system for content moderation or other high-stakes deployments.

3.1.2 Operational Requirements and Constraints

To remain operational in practice, the system must support batch inference for evaluation, handle missing OCR/context fields gracefully, and produce outputs in a consistent schema suitable for downstream analysis. Training and experimentation

are expected to require GPU acceleration for efficiency, while inference and reporting can be conducted on CPU for smaller workloads.

3.1.3 Reproducibility and Evaluation Requirements

We enforce reproducibility through versioned code, fixed random seeds, and a stable train/validation/test partition. Evaluation is designed to be robust under label imbalance by reporting macro-averaged metrics and examining per-class behavior. The prototype is considered complete when it consistently produces valid outputs on the held-out test set, reports stable macro-F1 and accuracy, and generates explanations that are readable and qualitatively aligned with evidence from OCR/context and the image.

3.2 Societal Impact

Memes have become a primary mode of communication on Bangladeshi social platforms, influencing political discourse, social commentary, and cultural identity. By enabling structured analysis of Bangla memes, our work contributes to low-resource multimodal language technology and supports broader research objectives such as studying narrative framing, polarization, and the spread of misinformation.

3.2.1 Potential Benefits

The system can reduce manual effort in meme analysis by producing structured role predictions and concise explanations, which may help researchers and analysts triage large volumes of content more efficiently. It also provides reusable infrastructure for future Bangla multimodal NLP research, where publicly available resources remain limited.

3.2.2 Potential Harms and Misuse

At the same time, this capability introduces societal risks if used without safeguards. Automated role labeling and explanation generation can be repurposed to amplify harassment, support political profiling, or produce misleading summaries that appear authoritative despite model uncertainty. Misclassification is particularly harmful in cases where a meme contains identifiable individuals, because incorrectly assigning a negative role may contribute to reputational harm or targeted abuse.

3.2.3 Mitigation and Responsible Use Boundaries

For these reasons, we present the system explicitly as a research prototype and emphasize that it should not be used as a stand-alone decision-maker in moderation or other high-stakes settings. Any real-world deployment would require human oversight, transparent reporting of limitations, and careful validation across culturally nuanced cases such as sarcasm, irony, and context-dependent humor.

3.3 Environmental Impact

Developing a multimodal learning system requires compute for training, embedding extraction, and repeated experimentation, which carries an associated energy cost.

3.3.1 Compute Footprint and Key Drivers

The dominant computational cost in this work arises from training and representation extraction, while inference is comparatively lightweight once the system is trained. The footprint is influenced by the number of training runs, the size of the pretrained encoders, and the extent of hyperparameter exploration.

3.3.2 Sustainability Measures

We reduce environmental impact by building on pretrained language and vision encoders rather than training from scratch, limiting exhaustive hyperparameter sweeps, and caching intermediate representations (e.g., precomputed embeddings) so that repeated evaluation does not require re-running expensive extraction pipelines. We also adopt efficiency-oriented practices such as checkpointing and early stopping to avoid unnecessary compute. As future work, we plan to further reduce compute by exploring smaller encoders and parameter-efficient adaptation methods, and by reporting compute budgets (e.g., GPU-hours) alongside performance to encourage more sustainable experimentation.

3.4 Ethical Issues

This project operates at the intersection of user-generated content, sensitive social discourse, and automated interpretation, which naturally raises ethical concerns.

3.4.1 Data Ethics: Consent, Privacy, and Copyright

Although memes were collected from public social media sources, public accessibility does not automatically imply informed consent for research reuse. Memes may also include personal identifiers (names, faces, phone numbers) embedded in images or text, raising privacy concerns. In addition, memes are often derivative works, and redistribution may introduce copyright and takedown obligations.

3.4.2 Model Ethics: Bias, Harmful Content, and Over-Reliance

Meme content can contain political hostility, harassment, or discriminatory language, and models trained on such data may learn undesirable associations or reflect biases present in the source distribution. Explanations can further increase the risk of over-reliance because they may appear authoritative even when the model is uncertain or when the meme relies on cultural context that is difficult to capture computationally.

3.4.3 Mitigation Practices and Professional Responsibilities

To address these concerns, we treat the dataset and outputs as research artifacts, recommend restricting redistribution, and encourage removing or masking personal identifiers where feasible. We also recommend annotator safety practices (clear warnings, the ability to opt out, and limiting exposure to disturbing material) and explicitly document model limitations and appropriate usage boundaries so the system is not presented as ready for automated moderation or other consequential applications.

3.5 Standards

Although this thesis does not claim formal compliance with a specific regulatory or industry standard, the development and reporting practices are aligned with widely accepted documentation norms in responsible machine learning.

3.5.1 Documentation and Reporting Practices

We structure dataset reporting around established dataset documentation principles (collection sources, annotation process, intended use, and limitations) and describe the trained system using model-reporting conventions that emphasize evaluation methodology, known failure modes, and appropriate usage.

3.5.2 Reproducibility Norms

We follow standard experimentation norms by maintaining clear train/validation/test splits, fixed random seeds, versioned code, and logged configurations, supporting auditability and repeatable results.

3.6 Project Management Plan

The project progressed through a sequence of phases that reflect the lifecycle of an applied machine learning system: building the dataset, creating reliable annotations, preparing multimodal inputs, training and integrating models, and finally evaluating and documenting findings.

3.6.1 Work Packages and Execution Narrative

We began by collecting memes from public sources, organizing them into a consistent structure, and removing duplicates and unusable instances. Because annotation was a primary bottleneck, we developed and iteratively improved an annotation tool and guideline set, enabling faster labeling and more consistent decisions across contributors. After the dataset stabilized, we prepared multimodal representations by extracting OCR text, incorporating contextual fields, and standardizing image inputs. With these components in place, we trained a role predictor and integrated modalities through a fusion strategy designed to leverage visual evidence without becoming overly sensitive to noisy text or sparse context. The final stage consisted of quantitative evaluation (macro-F1 and accuracy) and qualitative inspection of

model behavior and generated explanations, followed by iterative refinement of the write-up.

3.6.2 Schedule (High-Level)

At a high level, the schedule is organized around dataset freeze milestones (collection and annotation completion), baseline modeling completion, multimodal fusion completion, final evaluation, and report finalization. This milestone-driven schedule ensures that modeling and evaluation are conducted on a stable dataset version, and that analysis is not repeatedly invalidated by ongoing changes to the underlying data.

3.6.3 Resource Management

Resource planning is captured through the allocation of team responsibilities (collection, annotation supervision, model development, and reporting) and the specification of required tooling and compute. GPU-enabled environments are prioritized for training and embedding extraction, while CPU environments are sufficient for preprocessing, ablations on smaller subsets, and evaluation/report generation. Version control and experiment logging are used to coordinate work and preserve reproducibility. Budget estimation is intentionally omitted at this stage.

3.7 Risk Management

The risks in this project arise from the nature of memes as highly reusable templates, from the subjectivity of role interpretation, and from the sensitivity of real-world social content.

3.7.1 Technical and Data Risks

A central technical concern is that memes are frequently reposted with minor edits, which can unintentionally introduce near-duplicates across training and testing splits and inflate evaluation results. To preserve the integrity of our conclusions, we treat leakage prevention as a first-class requirement by de-duplicating and adopting split strategies that keep visually or semantically equivalent memes within a single partition. Another recurring risk is that the role distribution is naturally imbalanced; without corrective measures, models tend to over-predict dominant roles and appear accurate while failing on minority cases. We therefore rely on imbalance-aware training and report macro-averaged metrics and per-class behavior so that improvements reflect general performance rather than majority-class bias. Beyond labeling, the quality of OCR and context fields introduces additional uncertainty because Bangla memes often contain mixed-language text, stylistic spelling, and visual distortions, which can lead to incomplete or noisy extractions and unstable predictions. Our pipeline mitigates this by normalizing text, constraining excessively long inputs, and ensuring the system continues to operate when optional fields are missing.

3.7.2 Annotation and Operational Risks

We also recognize that role annotation is inherently subjective and culturally dependent; to limit label noise, we follow written guidelines, use pilot rounds to calibrate decisions, and apply periodic audits and adjudication for ambiguous cases. Operational risks such as limited GPU availability and training instability are reduced through embedding caching, checkpointing, and early stopping so that experiments remain feasible under realistic resource constraints.

3.7.3 Ethical and Misuse Risks

Ethical risks remain important throughout: memes can contain harmful or politically sensitive content, which can affect annotators and can be learned by the model in undesirable ways, and the system may be misused for profiling or harassment if deployed without oversight. For these reasons, we frame the work as research-oriented, recommend human-in-the-loop usage for any downstream setting, and encourage privacy-aware handling of data (including restricting redistribution and masking identifiers where feasible).

3.8 Economic Analysis

This section provides a high-level cost–benefit discussion for developing and maintaining the proposed meme understanding prototype.

3.8.1 Cost Drivers

The primary investment in this work is front-loaded: time spent collecting and curating data, effort spent creating consistent annotations, and compute consumed during training and embedding extraction. Annotation typically dominates costs because it requires human judgment for culturally nuanced content, and the development of an annotation tool is justified as a one-time investment that reduces per-item labeling time and improves consistency. Compute costs are comparatively easier to control because training is performed intermittently and the pipeline can reuse cached representations for repeated experimentation, while inference is lightweight once the system is trained.

3.8.2 Benefits and Value Proposition

On the benefit side, the prototype can reduce manual effort in analysis workflows by providing structured role predictions and concise explanations, enabling researchers or analysts to triage large meme volumes more quickly and to focus attention on complex or ambiguous cases. Over time, as the dataset expands and the pipeline stabilizes, incremental updates become less expensive than the initial build, making the system increasingly cost-effective for research and monitoring use cases.

3.8.3 Assumptions and Limitations

Any budgeted version of this analysis should state assumptions explicitly (annotation time per instance, compute platform pricing, and expected reduction in manual

review time) to ensure estimates remain transparent and credible. We also emphasize that economic value depends on integration into a workflow where outputs are verified by humans, particularly for sensitive or high-impact content.

Chapter 4

Dataset and Proposed Methodology

4.1 Dataset Preparation

4.1.1 Motivation and Existing Resources

Although a few Bangla meme datasets have been released in prior work, they are primarily designed for coarse-grained meme classification (e.g., category, sentiment, or offensive content detection). None of the available resources were compatible with our target task, which requires entity-level role annotations with supporting explanations and contextual metadata. Therefore, we constructed a new dataset tailored to our problem setting and cite the most relevant existing datasets for reference.

4.1.2 Data Collection from Social Media Sources

We collected Bangla memes from publicly accessible Facebook pages and groups that regularly post meme content across diverse themes. Our sources included active meme communities and topic-focused groups/pages such as *Rangtage*, *Political Meme Posting*, and 12 additional public pages/groups. This selection was intended to maximize diversity in meme templates, language style, and topical coverage.

4.1.3 Meme Extraction Pipeline

For most sources, we implemented an automated extraction pipeline using a custom Selenium-based script to navigate posts and download images along with their corresponding post URLs. A smaller portion of the dataset was collected manually to handle cases where automated scraping was limited due to inconsistent page structure, media loading issues, or other access constraints. All collected memes were then organized into a consistent storage format and mapped to metadata records for downstream processing.

4.1.4 Annotation Tool and Labeling Procedure

To efficiently gather structured labels at scale, we developed a web-based annotation platform named **BMAT (Bangla Meme Annotation Tool)**. BMAT was

used to collect entity-level role labels and accompanying explanations, along with additional meme interpretation fields. We conducted a crowdsourcing-style annotation process involving students from multiple universities, including the University of Dhaka (DU), BRAC University, and RUET, while providing a small monetary incentive to participants. Using BMAT, we collected 2700+ annotations. In addition, the authors manually annotated 800 samples, resulting in 3500+ total collected annotations.

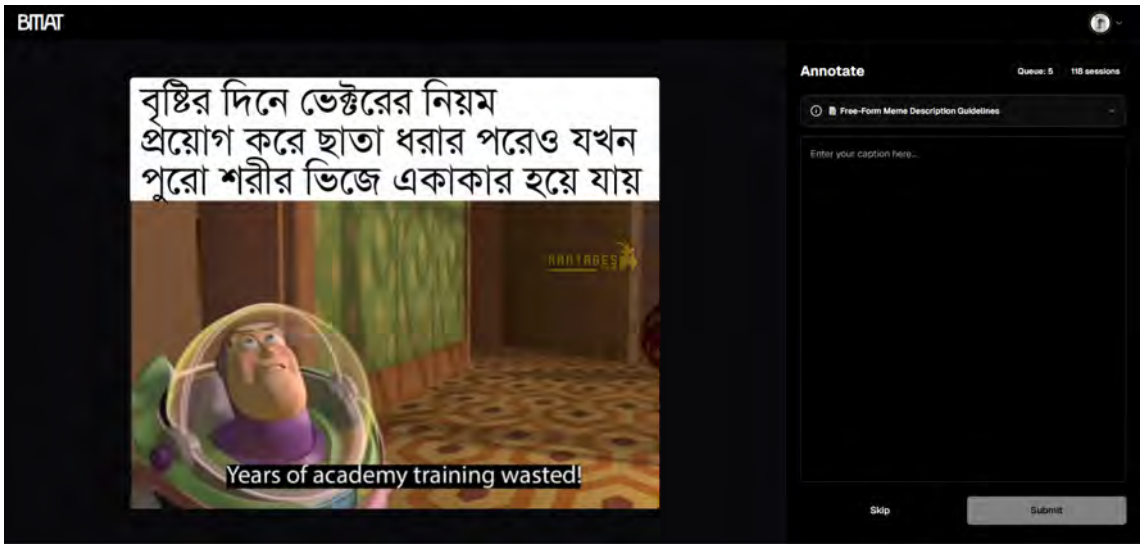

Figure 4.1: Annotation Website

4.1.5 Quality Control, Preprocessing, and Dataset Statistics

After annotation, we performed a review and cleaning stage to improve reliability and remove problematic samples. Importantly, all collected annotations were validated by the authors after the crowdsourcing phase to ensure completeness and consistency across label fields and explanations. This validation stage included checking whether each entry contained the required role and entity information, verifying that role assignments were coherent with the meme content, and flagging samples with unclear or contradictory interpretations.
Authors additionally performed preprocessing steps such as cross-checking entity and role fields, removing duplicates or unclear entries, and conducting OCR extraction and validation. After these procedures, we retained 2,658 high-quality annotated instances as the final dataset.

Table 4.1: Role label distribution in the final dataset, with role definitions.

Role	Count	Definition
Victim	1897	The meme shows this entity suffering, helpless, losing, or being treated unfairly.
Villain	1560	The meme blames, criticizes, mocks, or portrays this entity as the cause of problems or "bad."
Hero	625	The meme glorifies, praises, supports, or portrays this entity as "winning" or doing something right.

4.1.6 Final Dataset Schema

Each instance contains both annotation fields and supporting metadata to enable multimodal modeling and analysis. The final dataset includes the following columns: `entity_1`, `role_1`, `role_explanation_1`, `entity_2`, `role_2`, `role_explanation_2`, `humor_explanation`, `context`, `domain`, `OCR`, `image_description`.

Entities and role assignment. In our dataset, an *entity* refers to a specific real-world target that the meme is "about" or explicitly frames as an actor in the narrative. Entities can be people (e.g., politicians, celebrities), organizations (e.g., parties, institutions), groups (e.g., "students", "government"), or sometimes a country or community when the meme clearly treats it as a single participant.
For each selected entity, annotators assign one role label from {*Hero, Villain, Victim*} based on how the meme positions that entity. The *Hero* role is used when the meme praises, supports, glorifies, or frames the entity as "right" or "winning." The *Villain* role is used when the meme blames, criticizes, mocks, or presents the entity as responsible for harm or as "bad." The *Victim* role is used when the meme depicts the entity as suffering, losing, helpless, or being treated unfairly. These roles were chosen because they capture the most common narrative framing patterns in political memes and provide a compact, interpretable set of labels for supervised modeling.
Entity selection follows a simple principle: choose the entity (or two entities) that the meme most strongly targets through its visual cues, OCR text, and implied context. When multiple names or groups appear, priority is given to the entity that receives the strongest sentiment or narrative framing (praise/blame/suffering) and is central to the meme's message. This design ensures that the role labels reflect the meme's primary argumentative focus rather than peripheral mentions.

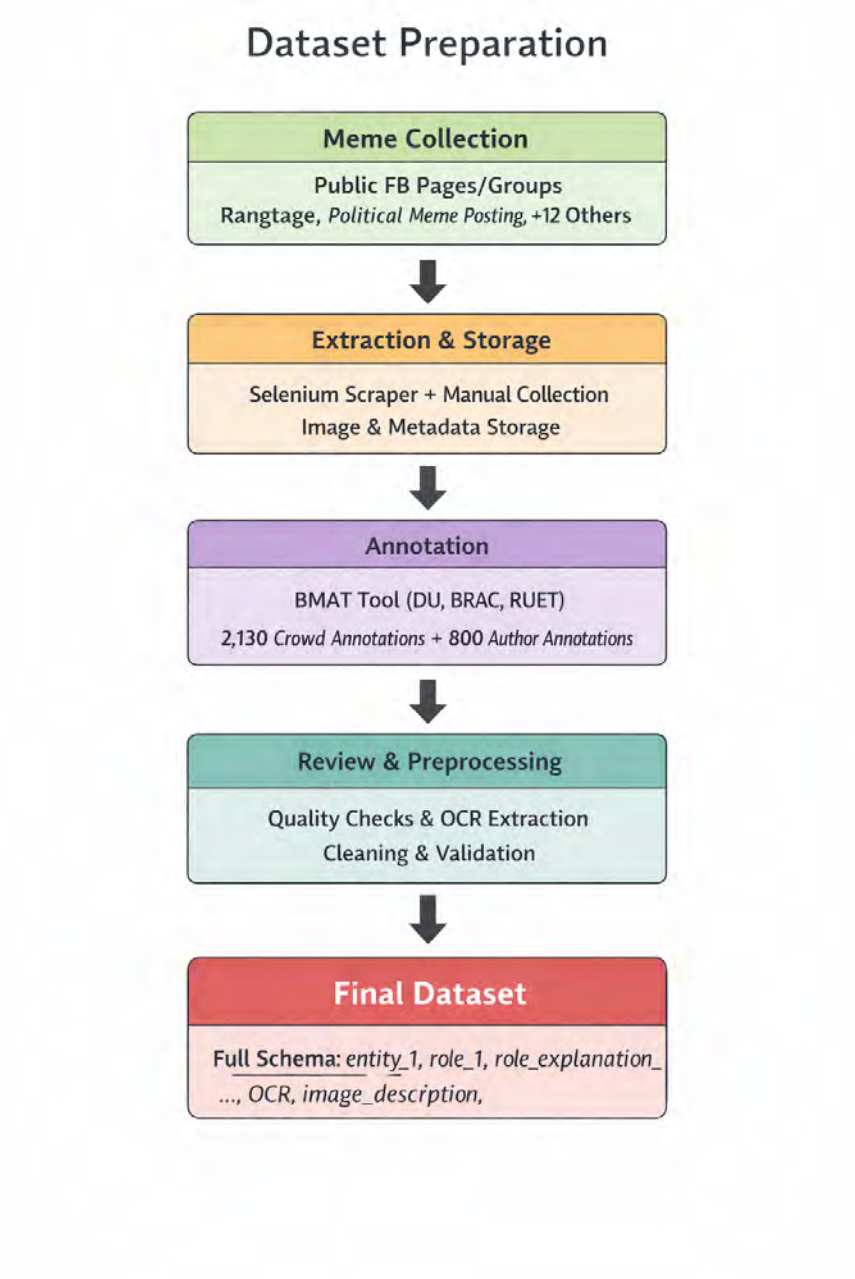

Figure 4.2: Dataset schema illustration

4.2 Design Process or Methodology Overview

Our thesis work is implemented through three different pipelines: two different alternative approaches for entity role classification and one for meme humor explanation generation. For all our pipelines, the meme is treated as a multimodal unit where meaning is constructed through the interaction of the visual scene, the text embedded within it detected through OCR, and the context. Additionally, our formulation is entity-centric. This means that rather than processing the entire meme and assigning it a label, we explicitly process the entity since the same meme may contain different entities with different narrative roles.

One of the most important methodological decisions that we took for our work was to implement two role classification pipelines. First, we implemented a controlled

and modular baseline where explicit processing and fusion of the different modality components through the use of lightweight attention is used. Although this is an interpretable and computationally efficient approach, the accuracy plateaued at around 0.70. This is why we implemented a second approach where a vision language model is used for more end-to-end multimodal processing. This is important since it will allow us to assess whether the errors that the explicit approach cannot handle can be mitigated through the use of end-to-end multimodal processing.

4.2.1 Triple-Stream Role Classification (Vision + OCR + Knowledge Fusion)

This pipeline utilizes a supervised multimodal classifier to predict one of three possible roles (victim, villain, hero) for each meme instance, utilizing three complementary input modalities: the meme image, the OCR text from the meme, and a context information relevant to the meme. We start off by loading the dataset and filtering the dataset by dropping off rows with empty OCR and context value, as both of these are required inputs for the architecture. We then drop non-essential fields (second entity fields are not necessary for this pipeline), create local paths to the images, and perform integrity checks to verify that the referenced images actually exist.
To provide more context than is available from the OCR text alone, we generate the knowledge text modality by concatenating the entity name, the available context text, and the image description text into a single text sequence. We split the data into training and validation subsets, stratifying each to preserve class balance. When training, we utilize balanced sampling (WeightedRandomSampler) along with a class-weighted cross entropy loss to prevent the model from collapsing to the most common role classes.
From an architectural point of view, this pipeline utilizes a triple-stream approach, where a pre-trained ViT encoder is used to generate visual token features from the meme image, BanglaBERT is used to generate contextual embeddings from the OCR text, and a multilingual MiniLM encoder is used to generate the context input modality. These three modalities are combined using a lightweight cross-attention mechanism, where the visual summary representation is used to query the combined text modalities (OCR and context), producing a single feature vector that is passed through a classifier to produce the final role prediction. This is a strong, interpretable baseline, as the contributions of each modality are made explicit, and systematic error analysis is straightforward (e.g., what is the contribution of the OCR text versus the image to the final prediction?).

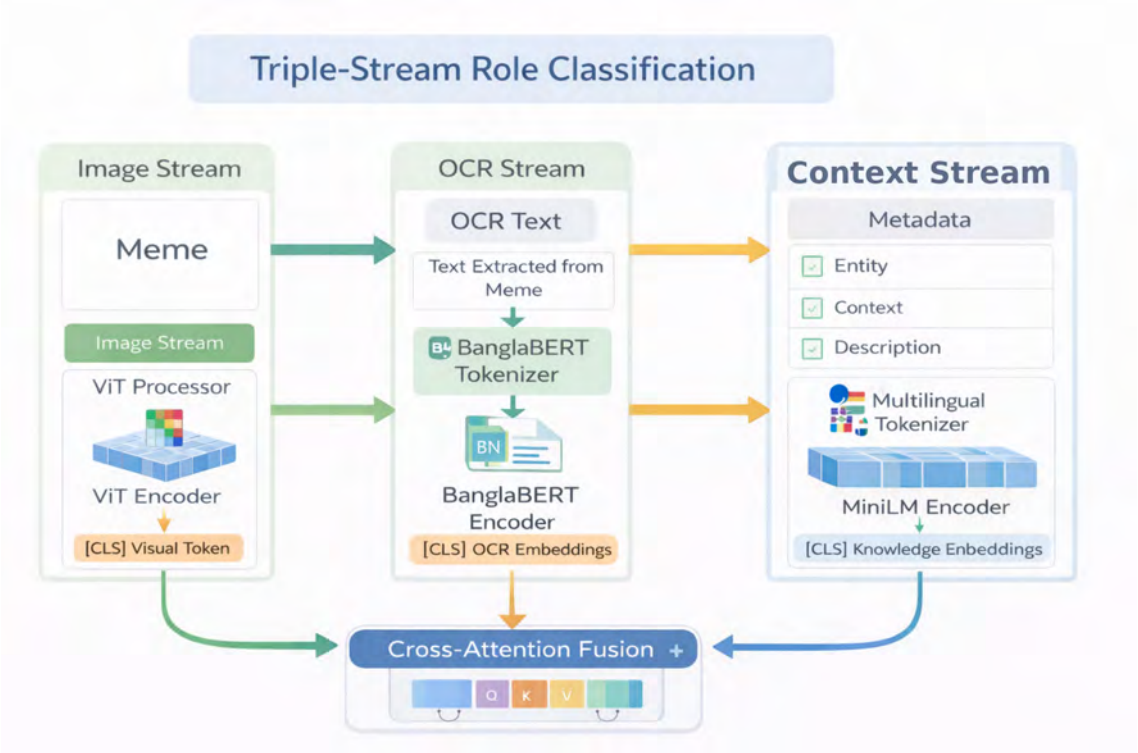

Figure 4.3: Triple-Stream Role Classification

4.2.2 Role Classification Using a Vision Language Model

While the initial pipeline explicitly fuses separately encoded modalities, memes often rely on implicit cues (sarcasm, visual framing, cultural references) that require more holistic cross-modal reasoning.[4] Therefore, we introduce a second role classification approach that uses a vision–language model and frames role prediction as an entity-focused reasoning task with structured outputs. The design process is as follows:

1. **Collect inputs for a meme instance:** For each meme, we gather the meme image, OCR text, and context (when available), along with the list of annotated entities that must be analyzed.

2. **Construct an entity-centric instruction:** We create an instruction that explicitly lists the entities and asks the model to assign a role to each one. This ensures the model stays anchored to the evaluation targets instead of generating a generic meme summary.

3. **Joint multimodal reasoning:** The model processes the image and text together, linking visual evidence (expressions, symbols, composition, emphasis) with OCR/context to infer the implied narrative role of each entity.

4. **Constrained role assignment:** For each entity, the model assigns exactly one role from {Hero, Villain, Victim}, reflecting whether the entity is framed positively, blamed/criticized, shown as suffering, or referenced without a strong role implication.

5. **Structured output for reliability:** The model is constrained to produce a machine-readable structured response containing one role decision per entity.

This reduces ambiguity, improves consistency across samples, and simplifies downstream analysis.

6. **Parsing and storage:** The structured output is parsed to extract role predictions, which are then stored for quantitative evaluation and error analysis.

Evaluation Factors

We evaluate role classification quality at the *entity level* using:

- **Overall accuracy** to measure the fraction of correct role predictions,
- **Macro-F1** as the primary metric to account for class imbalance,
- **Per-class precision/recall/F1** to identify difficult roles and systematic errors,
- **Confusion patterns** (confusion matrix) to understand recurring misclassifications,
- **Structured output validity** to quantify how often the output is parseable and includes exactly one valid role per entity.

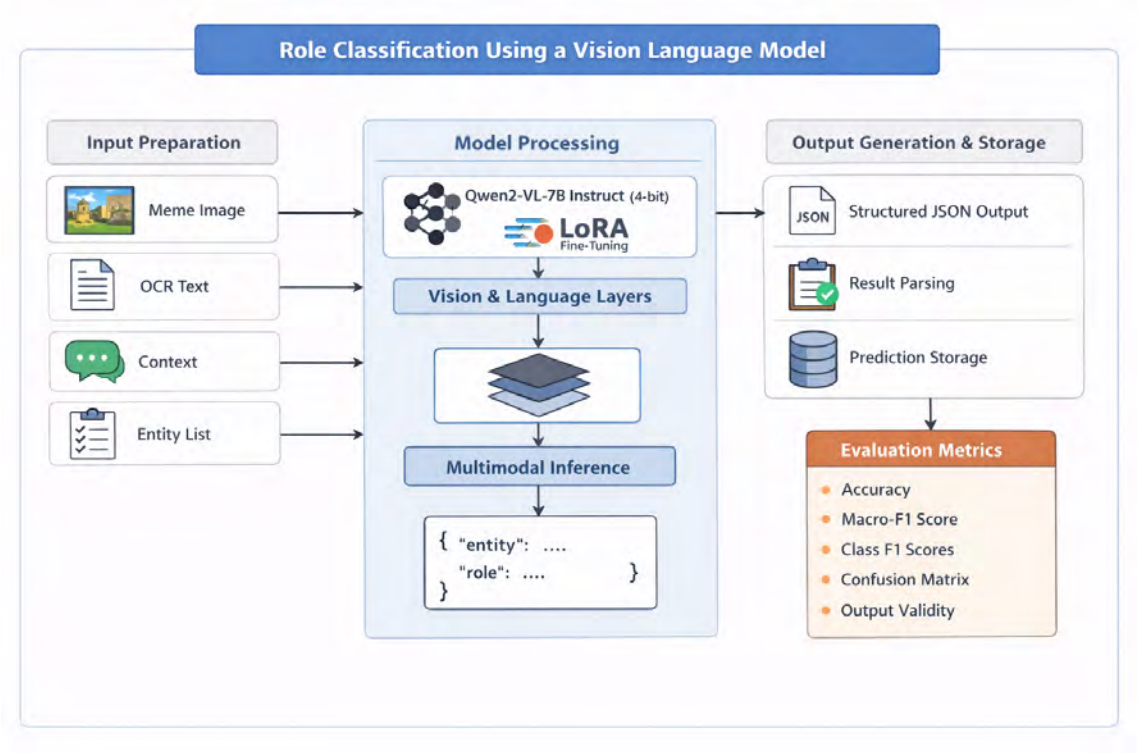

Figure 4.4: Role classification using a vision–language model and structured output parsing.

4.2.3 Meme Humor Explanation Generation

In addition to predicting entity roles, we aim to generate a grounded explanation of why a meme is humorous. This is treated as a multimodal generation task, where

the model must combine the visual scene with OCR/context and produce a short narrative explanation that reflects the meme's implied meaning rather than only restating literal text. The design process is as follows:

1. **Collect inputs for explanation:** For each meme, we use the image, OCR text, and context. When available, we also include the entity list and their roles so the explanation remains consistent about who is being referenced and who is being targeted.

2. **Grounded interpretation:** The model is instructed to interpret the meme as a complete artifact by linking visual details to OCR/context, since many memes are humorous due to contrast between what is shown and what is written.

3. **Explain the humor mechanism:** The model identifies what makes the meme funny (e.g., sarcasm, irony, exaggeration, parody, wordplay) and describes how this mechanism emerges from the image–text interaction.

4. **Generate a coherent narrative explanation:** The model produces a concise explanation that connects the setup (visual + textual evidence) to the implied punchline or social commentary, while avoiding unsupported claims.

5. **Normalize outputs for comparability:** We apply basic validation to ensure explanations are complete and readable, and remain grounded in the provided inputs.

Evaluation Factors:

For evaluation LLM-based judging was used over traditional automatic similarity metrics (e.g., BLEU, BERTScore)[5] [14] because humor explanations are *open-ended*: multiple explanations can be equally correct while sharing limited lexical overlap with a single reference [6]. In such cases, overlap-focused metrics can underscore valid paraphrases, and embedding-based similarity may still fail to verify key pragmatic requirements (e.g., identifying the correct humor target or mechanism). In contrast, an LLM-as-a-Judge can evaluate *semantic adequacy* under an explicit rubric by checking whether the explanation is grounded in the provided evidence, culturally plausible in the Bengali context, and whether it clearly articulates why the meme is funny in a coherent narrative. To reduce inconsistency and model-specific bias in scoring, we use a multi-model judge panel and aggregate their scores [24].
We evaluate explanation quality using rubric-based criteria designed for open-ended humor interpretation. Since multiple phrasings can accurately explain the same joke, we rely on structured judging rather than overlap-only scoring.

- **Rubric-Based Evaluation (LLM-as-a-Judge Panel):** We use an LLM-as-a-Judge setup where a panel of three models (`openai/gpt-4o`, `anthropic/claude-3.5-sonnet`, `google/gemini-3-flash-preview`) scores each generated explanation on a **1–10** scale using a fixed rubric [13]. We use **multiple judge models** because scoring can be sensitive to a single model's preference for wording, tone, or reasoning style; a panel reduces idiosyncratic

bias and yields a more stable estimate of quality by aggregating perspectives. Each judge receives the meme's *image description*, *context*, and *entities*, along with the model prediction and the reference humor explanation. The judge returns four dimension scores: `faithfulness`, `cultural_context`, `humor_deconstruction`, and `clarity`. We report the **average score across judges** for each dimension.

- **Visual Faithfulness / Groundedness:** Whether the explanation stays consistent with the described visual elements and provided context, and avoids introducing unsupported or hallucinated claims.
- **Cultural Context:** Whether the explanation captures Bengali cultural knowledge (e.g., local slang, political and social references, or shared community assumptions) instead of giving a generic reading.
- **Humor Deconstruction:** Whether the explanation clearly articulates *why* the meme is funny (e.g., sarcasm, irony, exaggeration, satire) and identifies the intended target of the joke when multiple entities are present.
- **Clarity and Coherence:** Whether the explanation is easy to read, logically structured, and understandable without requiring extra assumptions from the reader.

Final Score. For each generated explanation, we first compute the mean score for each dimension across the judge panel. We then compute an overall quality score as the **unweighted mean of the four dimension-wise averages**. This final score represents the panel's aggregate assessment of explanation quality, balancing (i) evidence-groundedness, (ii) culturally appropriate interpretation, (iii) correct decomposition of the humorous mechanism, and (iv) readability. Higher values indicate explanations that are simultaneously more faithful to the meme content, more culturally informed, more explicit about the humor, and clearer to a reader. ‘‘‘

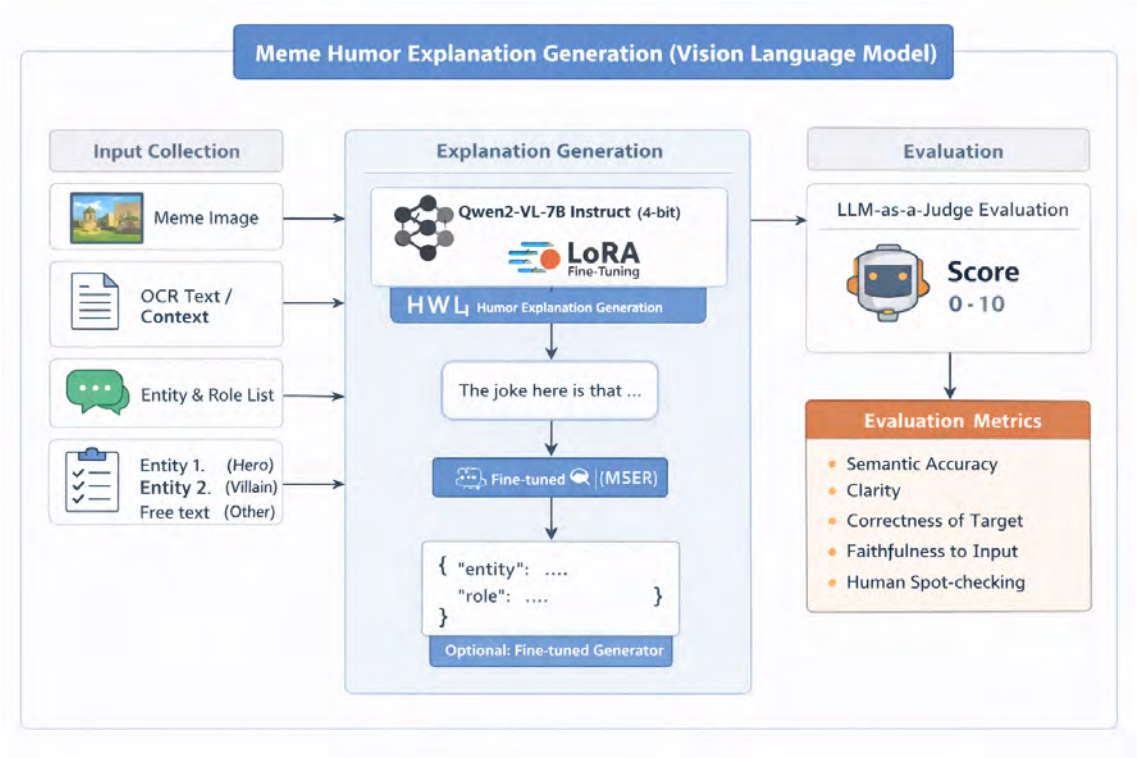

Figure 4.5: Meme Humor Explanation Generation

4.3 Model Specifications

This section documents the concrete model-level design choices for the two implemented pipelines. We describe the base model family, the input–output interfaces, the adaptation strategy, and the key training/inference configurations used in our experiments.

4.3.1 Role Classification (Triple-Stream Classifier)

This subsection specifies the model used in Triple-Stream Classifier, which performs entity role classification using three synchronized input streams: (i) the meme image, (ii) OCR text, and (iii) a constructed knowledge text block.

Task definition and label space

This pipeline is formulated as a supervised multi-class classification task. After filtering and final cleaning, the label space is restricted to three roles: *victim*, *villain*, and *hero*. Each sample is mapped to an integer label using the mapping victim $\rightarrow$ 0, villain $\rightarrow$ 1, hero $\rightarrow$ 2.

Input specification

For each sample, the model consumes:

- **Image stream:** a meme image loaded from local storage and processed using a ViT image processor (`google/vit-base-patch16-224-in21k`) to produce `pixel_values`.

- **OCR stream:** OCR text tokenized using BanglaBERT tokenizer (`csebuetnlp/banglabert`) with truncation/padding to a maximum length of 128 tokens.
- **Context stream:** a 'knowledge_input" sequence formed by concatenating *entity*, *context*, and *image_description* in a single text block, tokenized using a multilingual sentence-transformer tokenizer (`sentence-transformers/paraphrase-multilingual-MiniLM-L12-v2`) with truncation/padding to a maximum length of 256 tokens.

Backbone encoders

The architecture consists of three pretrained encoder backbones:

- **Vision encoder:** Vision Transformer, `ViTModel` (`google/vit-base-patch16-224-in21k`), producing a sequence of visual embeddings of hidden size 768. We choose this ViT backbone because memes often convey role cues through global visual structure (faces, symbols, emphasis, layout, and scene composition), and Vision Transformers are effective at capturing such frame-level semantics. ViT also provides token-level representations (including a CLS token), which is convenient for attention-based fusion with text features.
- **OCR encoder:** BanglaBERT, `AutoModel` (`csebuetnlp/banglabert`), producing token embeddings of hidden size 768. We choose BanglaBERT because OCR text in Bengali memes is predominantly Bengali and frequently includes code-mixed English tokens and informal spellings. A Bengali-focused pretrained encoder better preserves Bengali semantic nuance related to blame, praise, ridicule, or sympathy, which directly influences *hero/villain/victim* decisions. Although it is not optimized for English, subword tokenization enables reasonable handling of embedded English words, while English-heavy semantic framing is mainly captured by the separate context stream.
- **Context encoder:** Multilingual MiniLM, `AutoModel` (`sentence-transformers/paraphrase-multilingual-MiniLM-L12-v2`), producing token embeddings of hidden size 384. We choose MiniLM because the context and image-description fields are primarily English and serve as higher-level narrative framing for interpretation. MiniLM sentence-transformer models provide compact but semantically meaningful embeddings and are computationally efficient, which is important in a triple-stream architecture that already contains multiple pretrained backbones.

Fusion mechanism

Fusion is performed with a concatenation:

- The **query** is the visual CLS token (shape $[B, 1, 768]$).
- The **key/value** context is formed by concatenating the CLS tokens of OCR and knowledge streams along the feature dimension (shape $[B, 1, 1152]$).
- Linear projections map query and key/value into a common fusion dimension of 512, followed by multi-head attention (8 heads) and residual add–norm.

This produces a fused representation of shape $[B, 512]$ that is used for classification.

Classification head

The fused feature vector is passed through a lightweight MLP classifier:

$$512 \rightarrow 256 \rightarrow 3,$$

with ReLU activation and dropout of 0.3 between layers. The output logits correspond to the three role classes.

Training configuration

Training is done using a GPU with a stratified split of 70/15/15 for training/validation /testing to ensure that the class distribution is maintained across all splits of the dataset, thereby ensuring that the performance estimates are representative of the entire dataset. Class imbalance is handled by using two different methods: (i) WeightedRandomSampler, which increases the sampling rate of minority classes so that each mini-batch has a more balanced set of training instances, and (ii) class weighting of the cross-entropy loss, which heavily penalizes errors on minority classes, thereby preventing the model from learning to simply predict the majority class of each input role. We use AdamW with a weight decay of 0.01 and a learning rate of 2×10^{-5}, which is a common choice for fine-tuning transformer-based models, as it allows for sufficient adaptation to the target task while maintaining the stability of the pre-trained representations. A linear schedule with a warmup phase of 10% of the total training steps is used to gradually introduce updates from the start of training, which helps to prevent large gradient spikes that might destabilize the multimodal backbone of the model. The model is trained for 8 epochs, which is sufficient to expose it to the training data without risking excessive overfitting given the relatively modest size of our dataset. Gradient clipping with a norm of 1.0 is used to prevent gradients from exploding, which would destabilize training. The model checkpoint with the best validation accuracy is used to ensure that the final model has the best generalization performance, not necessarily the final training performance.

Inference and qualitative inspection

At inference time, predictions are obtained using an argmax over softmax-normalized logits. For qualitative inspection, the pipeline includes a visualization routine that displays a validation meme image together with the predicted role, true role, prediction confidence, and the corresponding OCR and knowledge text inputs to support interpretability and error analysis.

4.3.2 Role Classification (Vision–Language Role Labeler)

Base model and framework

We use a vision–language large model from the Qwen2-VL family in an efficient 4-bit configuration, loaded through the Unsloth `FastVisionModel` interface (`unsloth/Qwen2-VL-7B-Instruct-bnb-4bit`) with gradient checkpointing enabled for memory efficiency.

Input specification

Each training/inference example is structured as a chat-style message containing:

- **Image:** the meme image.
- **Text prompt:** an instruction block that defines the four roles (Hero, Villain, Victim, Other), requests reasoning per entity, and enforces a strict JSON-only output schema.
- **Context:** The textual context field.
- **Entity list:** one or two entities (e.g., `entity` and optionally `entity_2`) presented as bullet lines for analysis.

Output specification

The model is constrained to output **valid JSON only**, with a list of objects where each object includes: `entity`, `explanation` (reasoning), and `role` (one of Hero/Villain/Victim).

Supervision and data formatting

For supervised fine-tuning, we convert each dataset sample into a two-turn chat:

- **User turn:** image + instruction prompt + context + entity list.
- **Assistant turn:** the JSON object, populated from the dataset role label(s) and the corresponding role explanation field(s).

Adaptation method (LoRA)

We adapt the base vision–language model using parameter-efficient fine-tuning with LoRA [17], and we allow the adapters to update both the visual and textual reasoning pathways of the model. Concretely, LoRA adapters are enabled across the vision layers and language layers[40] , including the attention and MLP components, so the model can better align visual cues and textual evidence for role decisions without fully updating all base parameters. In our setup, we use a LoRA rank of $r = 16$ with $\alpha = 16$, set dropout to 0, do not train any bias terms, and fix the random seed to 3407 for reproducibility.

Training configuration

Training is carried out using TRL's `SFTTrainer`[18] together with Unsloth's vision-aware data collator, since our data is naturally represented as paired image–text examples in an instruction/chat format and the collator ensures consistent batching and padding for multimodal inputs. We use a per-device batch size of 2 and accumulate gradients over 4 steps to achieve a larger effective batch size without exceeding GPU memory limits, which is important when training a vision–language model with image tokens and long text sequences. We apply 10 warmup steps before the main optimization phase to gradually ramp up learning and avoid unstable early updates that could disrupt pretrained weights.

The model is trained for 2 epochs with a learning rate of 2×10^{-4}, which provides enough adaptation strength for LoRA-based fine-tuning while keeping training stable and limiting overfitting on a moderate-sized dataset. We use the `adamw_8bit` optimizer to reduce memory overhead and enable efficient training under limited hardware, and we apply a weight decay of 0.01 as regularization to improve generalization and reduce memorization of frequent patterns. A linear learning-rate schedule is used for its simplicity and stability, providing smooth decay across training. The maximum sequence length is set to 2048 tokens to accommodate the full instruction prompt together with OCR/context signals and to reduce truncation that could remove important evidence for role decisions. Finally, mixed-precision training (FP16 or BF16 depending on GPU support) is enabled to speed up training and reduce memory usage, making multimodal fine-tuning feasible while maintaining model quality.

Inference configuration

When we conduct inference, we create a new instruction prompt based on those we've used in the past (during training), in order to maintain the formatting of the output and encourage consistent behaviours. We also decode all tokens deterministically (as opposed to stochastically) and set our temperature to 0 (to avoid randomness) in order to create a smaller amount of variability across outputs, as well as to render all of our output valid JSON. Specifically, we have disabled sampling and allowed the generation of up to 300 new tokens. The output generated will be parsed as valid JSON, and we will extract our predictions of role for each of the entities in the output using the results of the parsing. Any outputs that do not parse as valid JSON will be flagged as invalid; if there is any instance of an entity that was not returned as part of the response, it will be explicitly handled so that the evaluation will have a well-defined outcome.

4.3.3 Humor Explanation Generation

Base model and framework

This pipeline uses `unsloth/Qwen2-VL-7B-Instruct-bnb-4bit` [39] via Unsloth `FastVisionModel` (4-bit loading, gradient checkpointing).

Input specification

For each meme, the generator receives:

- **Image:** the meme image.
- **Context and OCR text:** included verbatim when available.
- **Entities and roles:** a short list that explicitly states entity names and their role labels (e.g., `- entity (role)`).

The full prompt is assembled as a chat message, where the system prompt frames the task as explaining *why* the meme is funny, conditioned on the provided context and entity-role information.

Output specification

The model outputs a single natural-language humor explanation text. During baseline generation, decoding is deterministic to reduce variance across runs:

- max new tokens = 128
- `do_sample = False`

Fine-tuned generator (LoRA, optional variant)

In addition to baseline prompting, we also define a supervised fine-tuning variant that trains the model to reproduce the dataset's ground-truth `humor_explanation` . The training message format mirrors the inference structure, but the assistant target is the gold explanation text.
When the LoRA-based variant is used, we apply the same LoRA setup style (tuning vision + language + attention + MLP; $r = 16$, $\alpha = 16$, dropout = 0, bias none, seed 3407).

Evaluation model: LLM-as-a-Judge (multimodal panel)

Because humor explanation quality is difficult to capture with a single lexical metric, we evaluate generated explanations using a **panel** of external vision-capable judge models. Each judge independently reads:

- the meme image (via its image description in the input),
- the meme's context and OCR,
- the entity/role list,
- the ground-truth explanation,
- and the model prediction,

and then outputs an **individual score** according to the rubric dimensions (`faithfulness`, `cultural_context`, `humor_deconstruction`, `clarity`) on a **1–10 scale**, along with a short justification. We report scores per judge model and also use the mean across judges for a more stable aggregate estimate of explanation quality.

Chapter 5

Result Analysis

5.1 Performance Evaluation

In this section, the two role-classification pipelines and the humor explanation generation pipeline are tested on task-appropriate measures and testing processes. We present quantitative findings and afterwards interpret them to establish what the figures mean regarding the strengths and limitations of each design. The classification of roles is considered to be a multi-class problem that is being supervised and is an entity level problem. Humor explanation generation is evaluated using an LLM-as-a-Judge score [13], which reflects semantic alignment, groundedness, and overall explanation quality.

5.1.1 Evaluation Criteria

Role Classification. We report standard classification metrics: *accuracy* (overall correctness), *macro-averaged precision/recall/F1* to account for class imbalance by weighting each class equally, and *per-class precision/recall/F1* to identify role-specific strengths and weaknesses. For the vision–language model (VLM) pipeline, we also consider *output reliability* via structured-output validity (i.e., whether predictions are parseable and contain exactly one role per entity), since invalid structured outputs would reduce practical usability even if raw accuracy is high.
Explanation Generation. Because explanations are open-ended and may be correct even when phrased differently, we evaluate outputs using an *LLM-as-a-Judge* score on a 0–10 scale [13]. This score emphasizes whether the explanation captures the meme's implied meaning and humor mechanism, remains grounded in the image/OCR/context, and communicates a clear narrative that a reader can follow.

5.1.2 Testing Methods

For **Role Classification**, we evaluate on held-out splits and compute precision, recall, F1-score, and accuracy from the resulting predictions.Cross-attention classifier is evaluated on a test set containing 414 instances, while VLM-based role classification is evaluated on a held-out test set containing 414 instances. For **Humor Explanation Generation**, we generate an explanation for each meme in the evaluation split and score each output using the same judge prompt and rubric [13]; we then compare the average scores of the baseline and trained variants. Finally,

we perform spot-checking on a small subset of samples to verify that judge scoring aligns with human expectations, especially for culturally specific Bengali meme references.

5.1.3 Results and Reasoning

This subsection presents results for each pipeline and provides interpretation of what the results indicate about the designs.

Cross-Attention Classifier (Triple-Stream Fusion)

The cross-attention classifier was evaluated on a held-out validation set containing $N = 319$ samples. The model achieved an overall accuracy of **0.70**. Performance is stronger for *Victim* and *Villain* than for *Hero*: the F1-scores are 0.72 for Victim (support = 125), 0.73 for Villain (support = 121), and 0.58 for Hero (support = 73). This gap is consistent with two practical issues: (i) the Hero class has lower support, which reduces learning signals for class-specific patterns, and (ii) Hero is often more implicit in memes (e.g., the meme may mock one person while indirectly supporting another), making it harder to infer using modular features. The macro-average F1-score is **0.68**, reflecting the weaker minority-class performance, while the weighted-average F1-score is **0.69**, reflecting stronger performance on the more frequent classes.

Table 5.1: Classification report on the test set.

Class	**Precision**	**Recall**	**F1-score**	**Support**
Victim	0.71	0.73	0.72	185
Villain	0.70	0.77	0.73	165
Hero	0.67	0.52	0.58	64
accuracy		0.70		414
macro avg	0.69	0.67	0.68	414
weighted avg	0.69	0.70	0.69	414

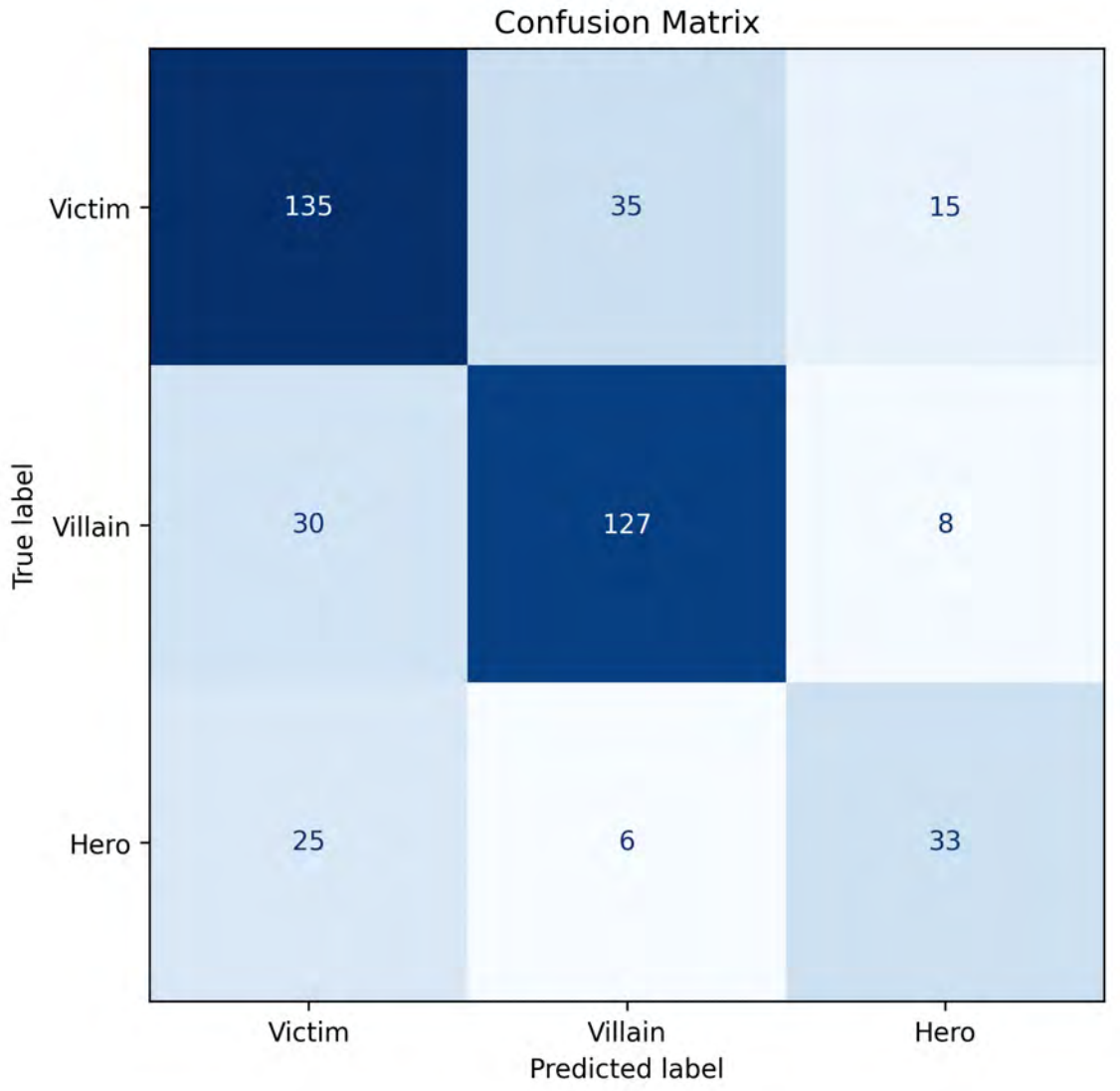

Figure 5.1: Confusion Matrix For Cross-Attention Classifier

Overall Accuracy. The model achieves an accuracy of **0.70** on the test set ($N = 414$).

Role Classification Using a Vision–Language Model

This pipeline is introduced to complement the explicit fusion approach by enabling holistic multimodal reasoning over memes, where meaning often emerges from implicit relationships between visual framing, textual tone, and shared cultural assumptions. While the triple-stream design models each modality separately and combines them through attention, many meme interpretations depend on narrative-level cues such as irony, social positioning, and the intended audience perspective, which are difficult to capture through feature fusion alone.

A vision–language model provides a unified representational space in which visual and textual evidence are processed jointly, allowing the model to align entities, actions, and implied intent across modalities. This formulation also supports *instruction-driven, entity-focused inference*, where the model can be guided to analyze specific targets and justify role assignments under a structured output

constraint. As a result, this pipeline emphasizes interpretability and semantic consistency, enabling direct integration with downstream explanation generation and qualitative analysis, in addition to quantitative role prediction.
On the held-out test set ($N = 414$), the pipeline achieves an overall accuracy of **0.8068**, which is substantially higher than Pipeline A. The macro-average F1-score is **0.7953**, indicating improved balance across roles compared to the baseline classifier. Class-wise, *Villain* and *Victim* achieve strong F1-scores (0.8182 and 0.8177 respectively), while *Hero* remains comparatively lower (0.7500), again consistent with its smaller support (64). Overall, these results support the rationale that end-to-end multimodal reasoning can reduce errors that remain difficult for modular fusion architectures.

Table 5.2: Classification report on the test set.

Class	**Precision**	**Recall**	**F1-score**	**Support**
hero	0.7083	0.7969	0.7500	64
villain	0.8811	0.7636	0.8182	165
victim	0.7889	0.8486	0.8177	185
accuracy		0.8068		414
macro avg	0.7928	0.8031	0.7953	414
weighted avg	0.8132	0.8068	0.8074	414

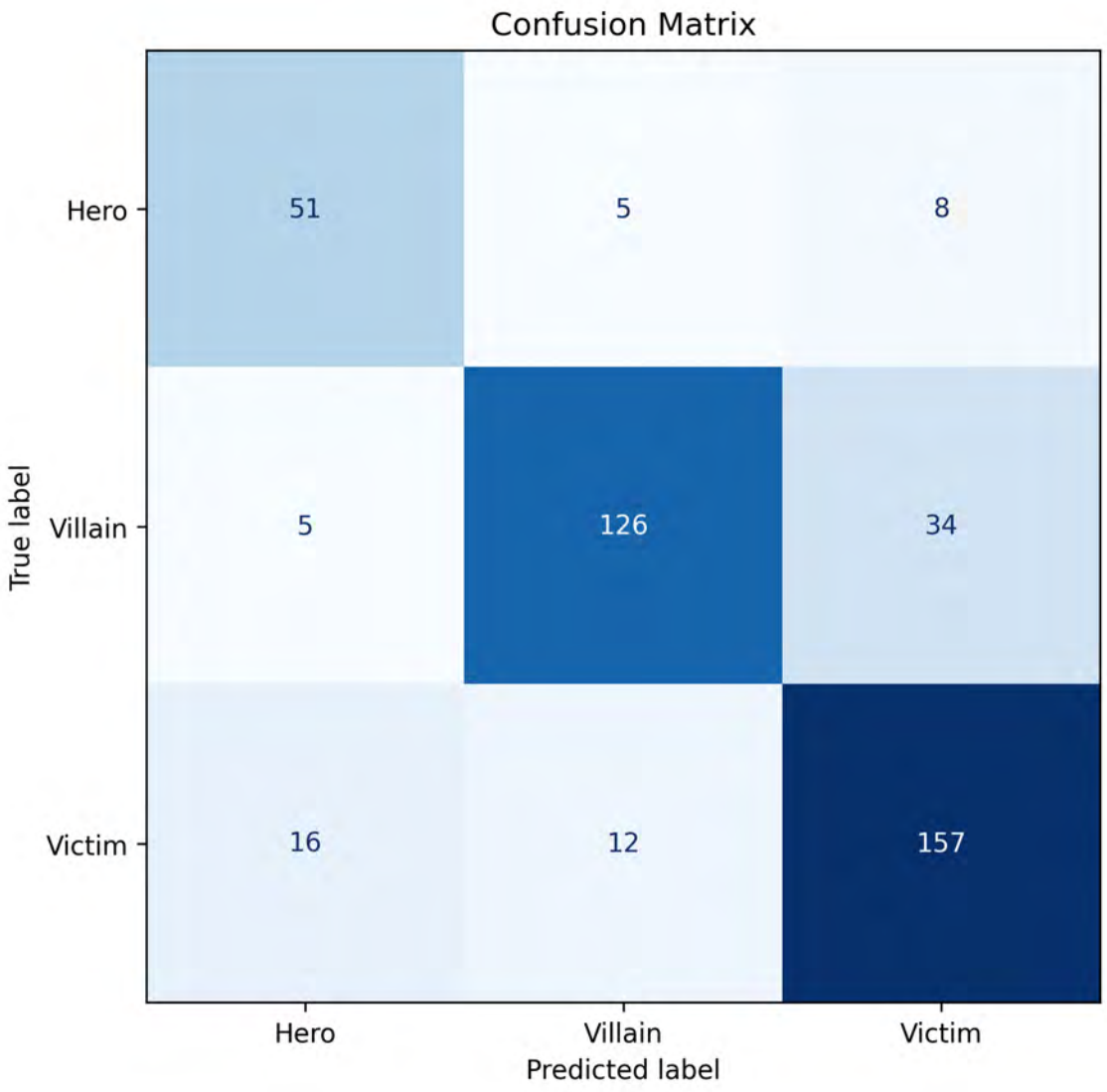

Figure 5.2: Confusion Matrix For VLM Classifier

Overall Accuracy: The model achieves an accuracy of **0.8068** on the test set ($N = 414$).

Meme Humor Explanation Generation

For humor explanation generation, we compare a baseline prompting approach with a trained/adapted variant. We evaluate explanation quality using an LLM-as-a-Judge panel [13], where three independent judge models (`openai/gpt-4o`, `anthropic/claude-3.5-sonnet`, `google/gemini-3-flash-preview`) score each prediction on a fixed rubric. The rubric contains four dimensions: **(i) faithfulness** (grounded in the described visual elements and provided context; no hallucinated details), **(ii) cultural_context** (Bengali-specific cultural knowledge such as slang, political/social references, and shared community cues), **(iii) humor_deconstruction** (explicitly explaining *why* the meme is funny and identifying the humor mechanism/target), and **(iv) clarity** (coherent, readable, and logically structured).
Each judge produces an *individual* 1–10 score for each dimension, and the overall score is computed as the unweighted mean of these four dimensions. Table 5.3

Judge	Variant	Faith.	Cult.	Humor	Clarity	Overall
`google/gemini-3-flash-preview`	Baseline	6.2	5.8	6.1	5.9	6.0
	Trained	7.9	7.4	7.8	7.3	7.6
`openai/gpt-4o`	Baseline	6.0	5.6	5.8	6.2	5.9
	Trained	7.3	6.8	7.2	7.1	7.1
`anthropic/claude-3.5-sonnet`	Baseline	6.4	5.9	6.1	6.4	6.2
	Trained	7.5	7.0	7.4	7.3	7.3

Table 5.3: LLM-as-a-Judge dimension scores (1–10) for humor explanation generation. Faith.=faithfulness, Cult.=cultural_context, Humor=humor_deconstruction. Overall is the unweighted mean of the four dimensions for each judge.

reports the baseline vs. trained/adapted dimension scores per judge. Overall, we observe consistent gains after training/adaptation: Gemini improves from **6.0** to **7.6** (**+1.6**), GPT-4o improves from **5.9** to **7.1** (**+1.2**), and Claude improves from **6.2** to **7.3** (**+1.1**). These improvements indicate that the trained/adapted generator produces explanations that are more consistently grounded in meme evidence (faithfulness), more culturally specific (cultural_context), more explicit in linking cues to the humorous mechanism and target (humor_deconstruction), and easier to follow as a short narrative (clarity). This judge-based evaluation is particularly appropriate because humor explanations are open-ended: correctness is primarily semantic and narrative, and multiple phrasings may accurately convey the same intended meaning [13].

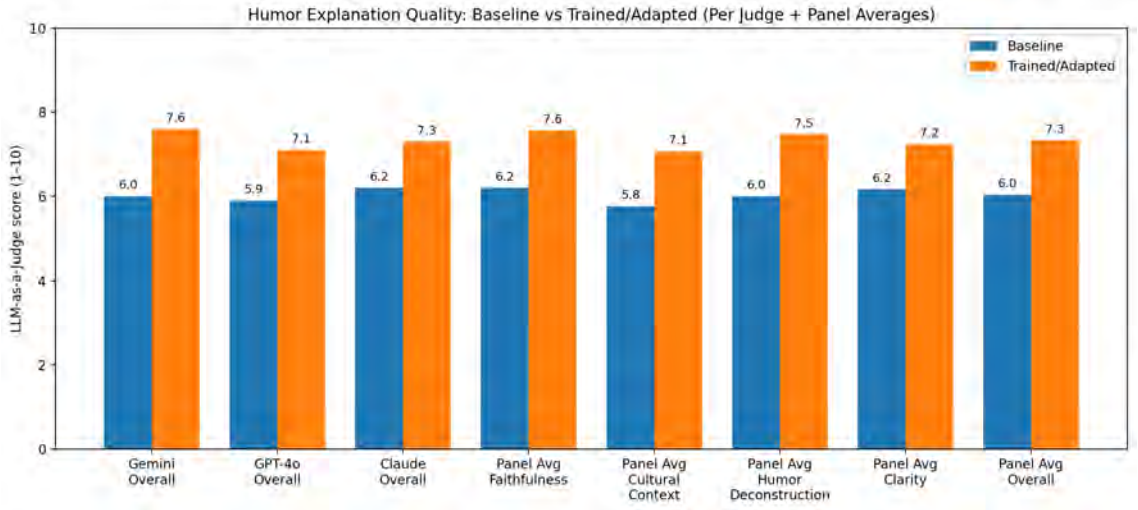

Figure 5.3: Explanation Score

5.2 Analysis of Design Solutions

This section analyzes the pipelines beyond raw performance, considering feasibility, efficiency, and practical trade-offs.

5.2.1 Role Classification: Cross-Attention Classifier vs. Vision Language Model Classifier

Performance trade-off. Cross-Attention Classifier provides a strong baseline with explicit control over each modality and achieves 0.70 accuracy. However, its weaker performance on the Hero class and the macro-F1 score indicate that modular encoders and lightweight fusion may not fully capture implicit narrative roles in complex memes. Vision Language Model Classifier improves accuracy to 0.8068 and macro-F1 to 0.7953, supporting the claim that joint reasoning in a vision–language model better handles subtle cross-modal cues.

Feasibility and efficiency. Cross-Attention Classifier is more computationally efficient for training and inference and is easier to interpret because each stream is explicit. Vision Language Model Classifier is more resource-intensive but produces higher accuracy and a structured output that supports reliable parsing and downstream analysis.

Strengths and weaknesses. Cross-Attention Classifier is strong for systematic debugging and controlled fusion, but can struggle when role inference depends on subtle intent or sarcasm. Vision Language Model Classifier is strong for such reasoning-heavy cases, but can be sensitive to prompt design and structured-output failures, requiring careful output validation.

5.2.2 Meme Humor Explanation Generation

Performance. Using the LLM-as-a-Judge rubric (faithfulness, cultural_context, humor_deconstruction, clarity), we observe consistent improvements after training/adaptation. The Gemini judge score increases from **6.0/10** (baseline) to **7.6/10** (trained). Under the same rubric, `openai/gpt-4o` improves from **5.9/10** to **7.1/10**, and `anthropic/claude-3.5-sonnet` improves from **6.2/10** to **7.3/10**. Averaging across the judge panel yields an overall increase from approximately **6.03/10** to **7.33/10**, indicating that training/adaptation improves explanation quality in a balanced way across groundedness, cultural specificity, explicit humor reasoning, and readability.

Feasibility and reliability. Prompt-only generation is fast and simple, while training improves consistency at additional compute cost. Judge-based scoring enables scalable evaluation, but it requires a stable rubric and prompt standardization to reduce inconsistency [13].

5.3 Final Design Adjustments

Based on the above evaluation, we make the following adjustments and recommendations.

Role classification. We prioritize improving the minority class (Hero) by increasing training examples and/or applying stronger imbalance handling (e.g., targeted sampling, class-aware weighting, or augmentation of hero-centric cases). We also refine role definitions and instructions to reduce ambiguity between borderline categories (e.g., Victim vs. Villain in political memes). For the VLM pipeline, we enforce stricter output validation and implement re-prompting or fallback handling for malformed or incomplete structured outputs.

Humor explanation generation. We strengthen grounding by explicitly instructing the generator to rely only on evidence present in the image/OCR/context, and we include entity/role information whenever available to keep the explanation aligned with the correct target. To improve judge reliability, we standardize the evaluation rubric and keep judge prompts fixed across all experiments. Finally, we use manual spot-checking on culturally sensitive subsets to identify recurring failure patterns and update prompts accordingly.

5.4 Statistical Analysis

We apply statistical techniques to quantify uncertainty and to test whether the observed improvements are meaningful rather than artifacts of sampling variation.
Role classification. For role classification, we report *bootstrap confidence intervals* (1,000 resamples) for accuracy and macro-F1 on the evaluation split to estimate variability under resampling of test instances. The best-performing role classification model (*Role Classification Using a Vision–Language Model*) achieves an accuracy of **0.8068** and a macro-F1 of **0.7953** on the held-out test set ($N = 414$). When comparing two role-classification variants on the same set of instances, we apply *McNemar's test* on paired predictions to determine whether differences in error patterns are statistically significant, since the predictions are dependent (paired at the instance level).
Humor explanation generation. Each meme receives paired evaluation scores (baseline vs. trained/adapted) on the same test items, so the comparison is paired rather than independent. Using the LLM-as-a-Judge panel, the mean overall score increases from **6.03/10** (baseline) to **7.33/10** (trained/adapted), i.e., **+1.30 points**. To test whether this improvement reflects a consistent shift across memes (and not only a few outliers), we apply the *paired Wilcoxon signed-rank test*, which is suitable for paired ordinal scores and does not assume normality. We additionally estimate uncertainty with *bootstrap confidence intervals* (1,000 resamples) for the mean judge score under each condition [13]. For completeness, we also report judge-specific paired comparisons (Gemini: **6.0→7.6**, GPT-4o: **5.9→7.1**, Claude: **6.2→7.3**), all computed on the same set of memes.

5.5 Comparisons and Relationships

This section compares the proposed approaches with each other and connects our design choices to broader trends in multimodal meme understanding. In particular, we contrast *Triple-Stream Role Classification (Vision + OCR + Knowledge Fusion)*, *Role Classification Using a Vision–Language Model*, and *Meme Humor Explanation Generation* as complementary components of a unified multimodal analysis framework.

5.5.1 Preliminary Model Comparisons on a US Politics Meme Dataset

Before finalizing the design of the *Triple-Stream Role Classification (Vision + OCR + Knowledge Fusion)* pipeline, we conducted an exploratory comparison on a US

politics meme dataset to evaluate how different multimodal model combinations perform for semantic role classification. This pilot study served as a model-selection phase rather than a final benchmark: it helped identify promising text–vision pairings and revealed recurring challenges, such as weak performance on minority or implicitly expressed roles. Since the US politics dataset differs from our target Bengali meme setting, we do not treat these results as final claims for our thesis task; instead, we use them to justify design decisions and motivate the need for stronger cross-modal reasoning.

Models	**Accuracy**	**Hero**			**Villain**			**Victim**			**Others**		
		Prec	**Rec**	**F1**	**Prec**	**Rec**	**F1**	**Prec**	**Rec**	**F1**	**Prec**	**Rec**	**F1**
BERT + EfficientNet	0.77	0.155	0.173	0.163	0.473	0.400	0.433	0.358	0.254	0.297	0.864	0.905	0.884
CLIP	0.74	0.078	0.034	0.045	0.409	0.513	0.469	0.398	0.276	0.321	0.856	0.843	0.849
BERT + ViT	0.78	0.155	0.171	0.162	0.530	0.680	0.590	0.398	0.276	0.321	0.820	0.860	0.830
BERT + ResNet	0.70	0.000	0.000	0.000	0.520	0.430	0.470	0.358	0.254	0.297	0.827	0.948	0.883

Table 5.4: Pilot comparison of multimodal model combinations for semantic role classification on a US politics meme dataset (used for architecture selection, not as the final benchmark for Bengali memes).

The pilot results highlight an important pattern: while overall accuracy remains reasonably high across multiple multimodal combinations, the *Hero* role is consistently difficult, showing very low F1-scores across most settings. This suggests that simple or loosely coupled multimodal fusion can struggle with roles that depend on implied intent, narrative framing, or subtle visual cues rather than explicit text. These findings motivated our decision to adopt an entity-centric formulation and a more structured fusion strategy in the triple-stream architecture, while also motivating the later shift to an end-to-end vision–language reasoning approach.

5.5.2 Comparison with Prior Work

We compare our role-classification results with the VECTOR model proposed by Sharma et al.[35], which predicts four roles (*Hero, Villain, Victim, Others*). Their class-wise scores show uneven behavior across the role categories: the *Others* class is predicted very strongly (F1 = 0.890), while the three semantically committed roles are substantially weaker (Hero F1 = 0.412, Villain F1 = 0.544, Victim F1 = 0.479). This pattern suggests that the model benefits from a broad catch-all category, but is less consistent when it must discriminate between nuanced narrative roles.
In contrast, our *Role Classification Using a Vision–Language Model* focuses on the three core roles (*Hero, Villain, Victim*) and exhibits more balanced behavior. The per-class F1-scores are close to each other—Hero (0.7500), Villain (0.8182), and Victim (0.8177)—and the corresponding precision/recall values remain relatively consistent across roles. This stability matters in meme understanding because real memes often frame roles implicitly; therefore, a practical system must reliably separate the role categories rather than performing well only on a default "Other" class.

Model	Roles	Hero (F1)	Villain (F1)	Victim (F1)
VECTOR [35]	Hero, Villain, Victim	0.412	0.544	0.479
CAT	Hero, Villain, Victim	0.28	0.53	0.49
Proposed VLM Classifier	Hero, Villain, Victim	0.7500	0.8182	0.8177

Table 5.5: Comparison between Sharma et al.[35] VECTOR , CAT and the proposed Vision–Language Role Classifier.

Beyond role prediction, our work expands the scope of Bengali meme understanding by adding a dedicated *Meme Humor Explanation Generation* pipeline. Existing Bengali meme datasets have largely focused on classification-style labels and typically do not provide both rich contextual fields and detailed natural-language explanations. Our dataset construction and pipeline design therefore extend prior work by jointly modeling (i) entity-centric role inference and (ii) grounded humor explanation generation using image, OCR, and context, enabling a more complete end-to-end understanding of Bengali memes rather than role prediction alone.

5.5.3 Triple-Stream Fusion vs. Vision–Language Model Reasoning

The *Triple-Stream Role Classification (Vision + OCR + Knowledge Fusion)* and the *Role Classification Using a Vision–Language Model* represent two complementary design philosophies for role inference. The triple-stream approach uses separate encoders for visual content, OCR text, and constructed knowledge, followed by an explicit fusion mechanism. This modular design supports controlled experimentation and detailed error analysis, as each modality's contribution can be inspected and improved independently. However, its weaker performance on nuanced or minority roles indicates that lightweight fusion may not fully capture deeper narrative dependencies present in complex memes.
In contrast, *Role Classification Using a Vision–Language Model* performs joint reasoning over image and text within a unified representational space and produces structured, entity-specific outputs. Empirically, it achieves higher overall performance in our Bengali meme setting, supporting the hypothesis that end-to-end multimodal reasoning is better suited for resolving implicit cues such as sarcasm, visual emphasis, and culturally grounded references.

5.5.4 Relationship Between Role Classification and Humor Explanation Generation

The *Meme Humor Explanation Generation* pipeline complements role classification by providing narrative grounding and interpretability. While role labels summarize how entities are framed (e.g., as hero, villain, or victim), explanations articulate why that framing arises and how the interaction between visual and textual elements produces humor. This relationship is particularly important in memes, where roles are often implied rather than explicitly stated. As a result, explanation generation serves both as an additional task and as a qualitative validation layer that helps interpret and contextualize the predictions made by the role classification models.

5.6 Discussions

This research's findings demonstrate the potential of memes as a truly multimodal and cultural item but also present some challenges to be overcome. We found that role classification is not only the detection of visual objects or matching of words in the text but also happens as part of interpreting the interaction of pictures, embedded text, and shared social context to convey an intention. The differing approaches to classifying roles in the Triple-Stream Model, which uses Vision + OCR (Optical Character Recognition) + Context Fusion, versus the Image Language Model, reveal a clear control and scalability difference. The high level of performance seen with the Triple-Stream Model suggests that the combination of learned features through lightweight fusion will create problems for rôles having elements of sarcasm, implied targets and, subtleties associated with framing/constructing a narrative. The substantially better performance of the Vision-Language Model supports the conclusion that performing end-to-end multimodal reasoning is a far better approach for finding these hidden cues.

Evidence for this interpretation comes from the way that generating an explanation for a particular behaviour tends to elevate the judge's score (as indicated by the table below) as the judge becomes better able to adjust to the dataset's linguistic style and cultural frame based on the generated explanation. The most important aspect of generating a legitimate explanation for the effectiveness of the meme and providing a brief explanation of how to interpret the meme will provide a level of interpretability of classifications for a particular category of meme that humans can use to determine the validity of the community's inference and hence whether there is any logical basis to the community's inference to be able to validate a meme based on the evidence available.

At the same time, the study reveals several limitations that shape how the results should be interpreted. The dataset reflects the biases of public social media sources and contains imbalanced role distributions, particularly for the *Hero* class, which contributes to weaker and less stable performance for that role. Annotation subjectivity further limits the upper bound on achievable accuracy, since different readers may reasonably disagree on how a meme frames an entity. In addition, variability in OCR quality and the absence of external contextual knowledge can obscure key signals needed for correct interpretation. On the evaluation side, the use of an LLM as a judge [9] provides a scalable way to compare explanations, but it remains an approximation of human judgment and must be treated as such.

Overall, this work suggests that effective meme understanding in low-resource and culturally specific settings requires models that combine joint multimodal reasoning with mechanisms for transparency and human validation. The results point toward a future in which stronger cross-modal fusion, richer contextual data, and more robust evaluation protocols can further improve both the accuracy and the trustworthiness of automated systems for analyzing socially and politically meaningful meme content.

Chapter 6

Conclusion

6.1 Key Findings

The thesis will be a multifaceted investigation of entity-focused role categorization and humor generation explanation of Bengali memes through multimodal learning. We show that it is crucial to consider memes as the combination of the image-text artifacts to capture its intended meaning and social functions. The suggested triple-stream cross-attention classifier is a solid and interpretable baseline, with a total accuracy of 0.70 on the test set, accentuating such challenges as the sensitivity of minority-class classifiers (especially when the role of the Hero is considered) and the inability to infer implicit narrative roles using only modular feature fusion.
Based on these findings, we propose a model of pipeline of vision-language multimodal reasoning with structured output. This method shows a higher level of performance with an accuracy of 0.8068 and macro-F1 score of 0.7953 on the held-out test set and proves that holistic cross-modal reasoning is more reasonable to interpret complex and ambiguous memes. To explain humor, we demonstrate that training the model leads to an increase in the average judge score by the trained model of 6.03/10 to 7.33/10, which means that the model is more semantically aligned and its explanations are more grounded. These findings combine with the previous findings to confirm the usefulness of entity-centric formulation, multimodal modeling, and scalable evaluation using LLM-as-a-Judge.

6.2 Contributions to the Field

It is a contribution to the understanding of multimodal memes comprehension that is currently being developed, especially concerning low resource and language specific languages like Bengali. First, we present an entity-based formulation of role classification, allowing to examine the representation of various entities within the same meme in finer detail, instead of interpreting such data into a single global category. Second, we present and comparatively analyze two complementary multi-modern pipelines a modular triple-stream union model and an end-to-end vision-language model which presents empirical results of the trade-offs among interpretability, computational efficiency, and classification performance.
Furthermore, this thesis offers a scalable procedure of assessing the open-ended humor clarifications with the help of an LLM-as-a-Judge structure according to structured prompts and statistical confirmation of this hypothesis. This methodology

helps to deal with one of the major weaknesses of conventional lexical measures and provides a feasible alternative to large-scale semantic assessment. Lastly, the role and humor explanation framework along with curated and annotated Bengali meme dataset becomes an excellent resource to be used in future studies of multilingual and culturally based multimodal comprehension.

6.3 Recommendations for Future Work

The findings of this thesis can be reinforced in a number of directions. One of the top priorities is the expansion of the dataset, especially the augmentation of the sample of underrepresented roles like the one of the Hero, so that the class imbalance could be minimized, and the strength would be enhanced. Secondly, a more rigorous validation protocol could be used in future iterations of the dataset: in this paper, the quality of labels and explanations was mostly checked by the authors, whereas it was not possible to fully validate them on a large scale. The next step would be to add (e.g., by introducing) specialist/domain-expert review (e.g., linguists, researchers in media studies, or experienced cultural annotators), of a representative subset of memes, with inter-annotator agreement reporting, to better measure the ambiguity, and to enhance the reliability of labels.
Adding more hints to the context like time and context and comment-driven information or social media metadata might further improve the capacity of the model to solve implicit references and culturally particular humor. In the modeling view, it can be further extended in the future through work on more aggressive cross-modal fusion mechanisms, like more multi-layer attention or contrastive pretraining specific to memes, to make visual and text representations more consistent. To generate explanations, it would be useful to incorporate human in the loop assessment or preference based learning to enhance faithfulness and decrease over generalization and generate a better quality evaluation signal than that produced by automated judging. Lastly, the broader of other languages with low resources and cross-cultural meme datasets would be a challenge that would test the generalizability and guide the development of more inclusive and multicultural multimodal AI systems.

Bibliography

[1] Md Tofael Ahmed et al. "Social media cyberbullying detection on political violence from Bangla texts using machine learning algorithm". In: *Journal of Intelligent Learning Systems and Applications* 15.4 (2023), pp. 108–122.

[2] David M Beskow, Sumeet Kumar, and Kathleen M Carley. "The evolution of political memes: Detecting and characterizing internet memes with multimodal deep learning". In: *Information Processing & Management* 57.2 (2020), p. 102170.

[3] Ali Braytee et al. "A Novel Dual-Pipeline based Attention Mechanism for Multimodal Social Sentiment Analysis". In: *Companion Proceedings of the ACM on Web Conference 2024*. 2024, pp. 1816–1822.

[4] Martin Juan José Bucher and Marco Martini. "Fine-tuned'small'LLMs (still) significantly outperform zero-shot generative AI models in text classification". In: *arXiv preprint arXiv:2406.08660* (2024).

[5] Chris Callison-Burch, Miles Osborne, and Philipp Koehn. "Re-evaluating the role of BLEU in machine translation research". In: *11th Conference of the European Chapter of the Association for Compuational Linguistics*. Association for Computational Linguistics. 2006, pp. 249–256.

[6] OECD STIP COMPASS. "AI RESPONDENTS FOR POLICY MONITORING: FROM DATA EXTRACTION TO AI-DRIVEN SURVEY RE". In: ().

[7] Mithun Das and Animesh Mukherjee. "Banglaabusememe: A dataset for bengali abusive meme classification". In: *arXiv preprint arXiv:2310.11748* (2023).

[8] Richard Dawkins. *The extended selfish gene*. Oxford University Press, 2016.

[9] Laura Dietz et al. "Principles and Guidelines for the Use of LLM Judges". In: *Proceedings of the 2025 International ACM SIGIR Conference on Innovative Concepts and Theories in Information Retrieval (ICTIR)*. 2025, pp. 218–229.

[10] Kazi Toufique Elahi et al. "Explainable Multimodal Sentiment Analysis on Bengali Memes". In: *2023 26th International Conference on Computer and Information Technology (ICCIT)*. IEEE. 2023, pp. 1–6.

[11] Ajeng Iva Dwi Febriana, Umaimah Wahid, and Alexander Seran. "The Distinction Among Political Humor Meme Creators in the Political Field of The 2019 Presidential Campaign for The Jokowi-Amin Pairing on Twitter". In: *Indonesian Journal of Social Science Research* 4.2 (2023), pp. 139–148.

[12] Shaik Fharook et al. "Are you a hero or a villain? A semantic role labelling approach for detecting harmful memes." In: *Proceedings of the workshop on combating online hostile posts in regional languages during emergency situations*. 2022, pp. 19–23.

[13] Jiawei Gu et al. “A survey on llm-as-a-judge”. In: *The Innovation* (2024).

[14] Michael Hanna and Ondřej Bojar. “A fine-grained analysis of BERTScore”. In: *Proceedings of the Sixth Conference on Machine Translation*. 2021, pp. 507–517.

[15] Ming Shan Hee, Wen-Haw Chong, and Roy Ka-Wei Lee. “Decoding the underlying meaning of multimodal hateful memes”. In: *arXiv preprint arXiv:2305.17678* (2023).

[16] Eftekhar Hossain et al. “Deciphering Hate: Identifying Hateful Memes and Their Targets”. In: *arXiv preprint arXiv:2403.10829* (2024).

[17] Edward J Hu et al. “Lora: Low-rank adaptation of large language models.” In: *ICLR* 1.2 (2022), p. 3.

[18] Shengchao Hu et al. “On transforming reinforcement learning with transformers: The development trajectory”. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 46.12 (2024), pp. 8580–8599.

[19] Md Maruful Islam et al. “Enhancing Offensive Bengali Social Media Meme Detection: A Weighted Ensemble Architecture for Predicting Type and Target Classes”. In: *2023 26th International Conference on Computer and Information Technology (ICCIT)*. IEEE. 2023, pp. 1–6.

[20] Nusratul Jannat et al. “An empirical framework for identifying sentiment from multimodal memes using fusion approach”. In: *2022 25th International Conference on Computer and Information Technology (ICCIT)*. IEEE. 2022, pp. 791–796.

[21] Saurav Joshi, Filip Ilievski, and Luca Luceri. “Contextualizing Internet Memes Across Social Media Platforms”. In: *Companion Proceedings of the ACM on Web Conference 2024*. 2024, pp. 1831–1840.

[22] Md Rezaul Karim et al. “Multimodal hate speech detection from bengali memes and texts”. In: *International Conference on Speech and Language Technologies for Low-resource Languages*. Springer. 2022, pp. 293–308.

[23] Akshi Kumar and Geetanjali Garg. “Sentiment analysis of multimodal twitter data”. In: *Multimedia Tools and Applications* 78 (2019), pp. 24103–24119.

[24] Haitao Li et al. “Llms-as-judges: a comprehensive survey on llm-based evaluation methods”. In: *arXiv preprint arXiv:2412.05579* (2024).

[25] Yang Liu et al. “G-eval: NLG evaluation using gpt-4 with better human alignment”. In: *arXiv preprint arXiv:2303.16634* (2023).

[26] Zhicheng Liu et al. “Ensemble Pretrained Models for Multimodal Sentiment Analysis using Textual and Video Data Fusion”. In: *Companion Proceedings of the ACM on Web Conference 2024*. 2024, pp. 1841–1848.

[27] Syrielle Montariol et al. “Fine-tuning and sampling strategies for multimodal role labeling of entities under class imbalance”. In: *Proceedings of the Workshop on Combating Online Hostile Posts in Regional Languages during Emergency Situations*. 2022, pp. 55–65.

[28] Levente Murgás et al. “Decoding memes: A comparative study of machine learning models for template identification”. In: *arXiv preprint arXiv:2408.08126* (2024).

[29] Abrar Shadman Mohammad Nahin et al. “Bengali Hateful Memes Detection: A Comprehensive Dataset and Deep Learning Approach”. In: *2024 International Conference on Advances in Computing, Communication, Electrical, and Smart Systems (iCACCESS)*. IEEE. 2024, pp. 01–06.

[30] Usman Naseem et al. “A multimodal framework for the identification of vaccine critical memes on Twitter”. In: *Proceedings of the Sixteenth ACM International Conference on Web Search and Data Mining*. 2023, pp. 706–714.

[31] Lynnette Hui Xian Ng, Adrian Xuan Wei Lim, and Roy Ka-Wei Lee. “Love-Hate Dataset: A Multi-Modal Multi-Platform Dataset Depicting Emotions in the 2023 Israel-Hamas War”. In: *Companion Proceedings of the ACM on Web Conference 2024*. 2024, pp. 1807–1815.

[32] Umera Wajeed Pasha. “Multilingual sexism detection in memes, a CLIP-enhanced machine learning approach”. In: *Working Notes of CLEF* (2024).

[33] Matthew E Peters et al. “Dissecting contextual word embeddings: Architecture and representation”. In: *arXiv preprint arXiv:1808.08949* (2018).

[34] Shraman Pramanick et al. “MOMENTA: A multimodal framework for detecting harmful memes and their targets”. In: *arXiv preprint arXiv:2109.05184* (2021).

[35] Shivam Sharma et al. “Characterizing the entities in harmful memes: Who is the hero, the villain, the victim?” In: *arXiv preprint arXiv:2301.11219* (2023).

[36] Shivam Sharma et al. “Findings of the CONSTRAINT 2022 shared task on detecting the hero, the villain, and the victim in memes”. In: *Proceedings of the Workshop on Combating Online Hostile Posts in Regional Languages during Emergency Situations*. 2022, pp. 1–11.

[37] Shivam Sharma et al. “MEMEX: Detecting Explanatory Evidence for Memes via Knowledge-Enriched Contextualization”. In: *arXiv preprint arXiv:2305.15913* (2023).

[38] Shivam Sharma et al. “What do you meme? generating explanations for visual semantic role labelling in memes”. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 37. 8. 2023, pp. 9763–9771.

[39] Peng Wang et al. “Qwen2-vl: Enhancing vision-language model’s perception of the world at any resolution”. In: *arXiv preprint arXiv:2409.12191* (2024).

[40] Maxime Zanella and Ismail Ben Ayed. “Low-rank few-shot adaptation of vision-language models”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2024, pp. 1593–1603.

[41] Yang Zhong and Bhiman Kumar Baghel. “Multimodal Understanding of Memes with Fair Explanations”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2024, pp. 2007–2017.

[42] Yi Zhou, Zhenhao Chen, and Huiyuan Yang. “Multimodal learning for hateful memes detection”. In: *2021 IEEE International conference on multimedia & expo workshops (ICMEW)*. IEEE. 2021, pp. 1–6.