BANMIME : Misogyny Detection with Metaphor Explanation on Bangla Memes

Md Ayon Mia[2,*], AKM Moshiur Rahman[1,*], Khadiza Sultana Sayma[3,*], Md Fahim[1,4,*,†], Md Tahmid Hasan Fuad[1], Muhammad Ibrahim Khan[3], AKM Mahbubur Rahman[1]

[1]*Center for Computational & Data Sciences* [2]*Dhaka International University*
[3]*Chittagong University of Engineering and Technology* [4]*Penta Global Limited*
[*]Equal Contribution [†]Project Lead
Correspondence: {mdayonrahman100, fahimcse381}@gmail.com

Abstract

Detecting misogyny in multimodal content remains a notable challenge, particularly in culturally conservative and low-resource contexts like Bangladesh. While existing research has explored hate speech and general meme classification, the nuanced identification of misogyny in Bangla memes, rich in metaphor, humor, and visual-textual interplay, remains severely underexplored. To address this gap, we introduce BANMIME , the first comprehensive Bangla misogynistic meme dataset comprising 2,000 culturally grounded samples where each meme includes misogyny labels, humor categories, metaphor localization, and detailed human-written explanations. We benchmark the various performances of open and closed-source vision-language models (VLMs) under zero-shot and prompt-based settings and evaluate their capacity for both classification and explanation generation. Furthermore, we systematically explore multiple fine-tuning strategies, including standard, data-augmented, and Chain-of-Thought (CoT) supervision. Our results demonstrate that CoT-based fine-tuning consistently enhances model performance, both in terms of accuracy and in generating meaningful explanations. We envision BANMIME as a foundational resource for advancing explainable multimodal moderation systems in low-resource and culturally sensitive settings. The code and dataset are publicly available at https://github.com/Ayon128/BANMIME.

Disclaimer: This paper contains elements that one might find offensive which cannot be avoided due to the nature of the work.

1 Introduction

In the digital era, internet and social media have fundamentally transformed global communication patterns, with approximately 5.56 billion internet users representing 67.9% of the global population[1] and 5.24 billion active social media participants worldwide[2] as of early 2025. Within this digital landscape, memes have emerged as powerful vehicles for entertainment and social commentary, blending visual and textual elements into easily shareable formats that now constitute a ubiquitous feature of online culture (Zannettou et al., 2018). Although their accessibility facilitates widespread adoption, this same characteristic enables the rapid spread of problematic content, particularly in culturally conservative regions such as Bangladesh, where women frequently become targets of degrading content. Consequently, the convergence of expanding digital connectivity with conservative so-

Figure 1: **Explainable misogyny classification via metaphor understanding.** Given a meme image and its associated text, the goal is to identify the underlying category of misogyny (e.g., shaming, objectification) and generate an explanation by interpreting the metaphorical meaning conveyed through the image–text interplay.

[1]https://www.meltwater.com/en/blog/digital-2025
[2]https://www.aiprm.com/technology-statistics/

Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 17825–17851
November 4-9, 2025

cial structures creates environments where misogynistic memes significantly impact women's participation and well-being. Previous research by Blake et al. (Blake et al., 2021) established clear correlations between online misogyny and offline violence against women, while (Paciello et al., 2021) demonstrated how exposure to sexist memes undermines moral reasoning and emotional processing. These findings collectively indicate that misogynistic memes extend beyond mere offensive content to constitute vectors for tangible societal harm. Moreover, (Kiela et al., 2020) identify unique moderation challenges posed by memes' multimodal nature, where harmful messaging emerges from the complex interaction between seemingly innocuous text and images.

Despite increasing research attention to this critical issue, substantial challenges persist in detecting misogynistic memes. Notably, (Fersini et al., 2022) developed a pioneering task on multimedia automatic misogyny identification, highlighting detection complexities across cultural contexts, while (Singh et al., 2023) documented persistent difficulties faced by contemporary models in identifying misogynistic content, particularly emphasizing the nuanced interplay between visual and textual elements. Nevertheless, the identification of misogynistic content in Bangla, a language with over 230 million speakers worldwide, remains significantly under-researched. Previous work on Bangla datasets has primarily focused on general meme identification or broad hate speech detection (Hossain et al., 2022b,a; Nahin et al., 2024); however, none specifically address the unique characteristics of misogyny in Bangla memes. To the best of our knowledge, we are the first to propose a comprehensive dataset, BANMIME (Bangla Misogynistic Memes), specifically designed to detect misogyny in Bangla memes by addressing this critical gap. Our dataset employs a fine-grained taxonomy of four categories capturing cultural nuances of misogyny: Stereotype, Objectification, Shaming, and Violence. Additionally, each meme is annotated with classification, humor labeling, metaphor localization, and comprehensive explanations that identify specific elements contributing to its misogynistic nature as illustrated in Figure 1. This approach follows similar reasoning to Hwang and Shwartz (Hwang and Shwartz, 2023), who introduced MemeCap for captioning and interpreting memes.

Our extensive experiments using both open-source and closed-source large vision-language models (LVLMs) reveal that while current models can achieve reasonable accuracy in detecting misogynistic content, they continue to struggle with generating accurate explanations that identify specific mechanisms through which misogyny is conveyed. This performance gap underscores the need for culturally-aware AI systems capable of understanding the contextual and multimodal nature of online misogyny. In light of these findings, our major contributions include:

- BANMIME the first comprehensive dataset of 2,000 Bangla memes annotated for misogyny detection with explanations and fine-grained categorization.
- A thorough benchmarking of open and closed-source large language models on misogyny detection and explanation generation tasks using zero-shot, Chain-of-Thought, and supervised fine-tuning approaches.
- Systematically explore various fine-tuning approaches, including standard, data-augmented, and Chain-of-Thought methodologies, revealing their relative efficacy for metaphor understanding and misogyny detection.

2 Related Work

Prior research on hate and misogyny detection has explored both unimodal and multimodal approaches, evolving from text-focused traditional models to transformer-based systems that integrate text and images.

2.1 Unimodal Approaches

Unimodal approaches typically process either text or image data. In the context of Bangla misogyny detection, Jahan et al. leveraged RNN, LSTM, and BanglaBERT-based embeddings, while Anzovino et al. applied SVM and Random Forests, both facing challenges with class imbalance (Jahan et al., 2023; Anzovino et al., 2018). Other studies explored TF-IDF features with Naive Bayes and GRU networks (Ishmam and Sharmin, 2019) on Facebook comments, as well as hybrid CNN–RNN architectures for aggression detection (Sadiq et al., 2021). Despite moderate success, unimodal systems often fail to capture implicit or context-dependent cues. (Haider et al., 2024) introduced a novel dataset and experimented with various state-of-the-art approaches for multilabel hate speech

identification in transliterated online social media texts.

2.2 Multimodal Approaches

To address the limitations of unimodal methods, multimodal approaches have emerged as a more effective solution, particularly for meme-based hate and misogyny. Several datasets have facilitated this research, including MUTE for Bangla memes (Hossain et al., 2022c), MultiOFF for offensive memes (Suryawanshi et al., 2020), MIMIC for multi-label misogyny (Singh et al., 2024), and MIMOSA for aggression in Bangla memes (Ahsan et al., 2024). These resources highlight challenges such as sarcasm detection, OCR errors, and imbalanced label distributions.

On the modeling side, both early and late fusion strategies have been explored, showing improvements over unimodal baselines, although fusion complexity often limits performance. Recent work emphasizes interpretable and reasoning-aware approaches, including attention- and graph-based methods (Rehman et al., 2025), debate-style reasoning between LLM-generated explanations (Lin et al., 2024), and brain-inspired fusion architectures that separate primary semantics from auxiliary context (Wu et al., 2024). CLIP-based methods, such as Hate-CLIP (Kumar and Nandakumar, 2022), explicitly model cross-modal interactions, capturing subtle contextual cues.

Cross-lingual and regional approaches have further extended detection beyond English. For instance, MuRIL and mBERT-based models have been applied to detect misogyny in Tamil and Malayalam memes (Mahesh et al., 2024), and ensemble-based methods have been proposed for multilingual robustness (Gu et al., 2022). These advances underscore the importance of robust multimodal and multilingual frameworks, while highlighting persistent challenges in explainability, cross-domain generalization, and low-resource languages such as Bangla.

3 BANMIME Dataset Collection and Annotation

The creation of BANMIME (Bangla Misogynistic Memes) involves a systematic collection and annotation process designed to address the critical gap in resources for detecting misogynistic content in low-resource languages. Our dataset comprises 2,000 carefully curated memes and employs a fine-grained taxonomy of four categories that reflects Bangla cultural contexts: Stereotype, Objectification, Shaming, and Violence. Figure 2 illustrates the overall pipeline of the BANMIME dataset development.

Data Collection We collected a total of 4,280 memes from predominant social media platforms utilized by Bangla speakers: the majority from Facebook (2,850 memes), a substantial portion from Instagram (950 memes), and a smaller number from Reddit (480 memes). This distribution reflects regional usage patterns where Facebook maintains higher penetration. Our data collection targeted public pages and communities between January 2020 and December 2024, employing temporally stratified sampling to ensure balanced representation across this period. Notably, meme creation surged during the COVID-19 pandemic, and our four-year span captures evolving themes from humorous and satirical to derogatory content, reflecting shifting gender-based stereotypes and debates.

Data Filtering We implemented a systematic filtering process to ensure dataset quality and relevance. Specifically, we removed: (1) non-readable memes with degraded image quality, (2) memes without textual content, (3) memes containing exclusively English text , and (4) directly nude pictures, as our study focuses on Bangla linguistic and cultural markers. This filtering process eliminated approximately 1,410 memes, leaving us with 2,870 memes with extractable, readable Bangla textual content that were retained for further analysis.

Data Cleaning During the cleaning phase, we applied deduplication algorithms to eliminate redundant instances of identical memes appearing across multiple sources. We systematically removed extraneous elements, including hyperlinks, URLs, and hashtags embedded in the text, as these were not central to the semantic analysis and could potentially introduce noise. This preprocessing removed an additional 870 duplicate entries and non-essential textual elements, resulting in a final curated dataset of 2,000 clean, text-bearing memes ready for annotation.

Data Annotation To annotate the BANMIME dataset properly, we recruited three native Bangla-speaking undergraduate annotators with expertise in Bangla meme culture and social media discourse for the annotation task. Their prior experience with meme creation and cultural context interpretation ensured high-quality

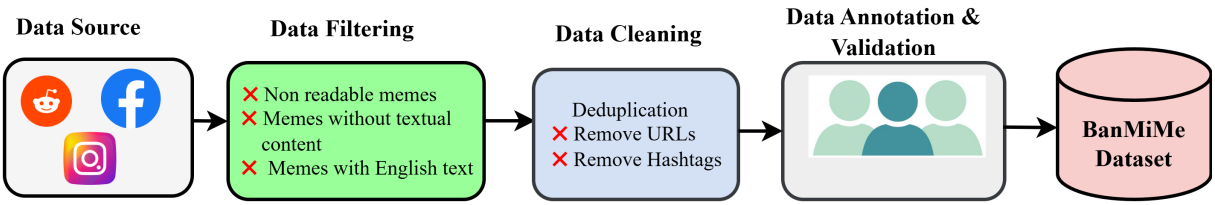


Figure 2: BANMIME dataset development pipeline illustrating the four-stage process: Data Source collection from social media platforms, Data Filtering to remove non-relevant content, Data Cleaning to eliminate duplicates and extraneous elements, and Data Annotation & Validation resulting in the final dataset.

annotations that captured nuanced misogynistic content. We provided comprehensive guidelines with annotators for the annotation process. The annotation guidelines are provided in the Appendix Section A. The annotators were instructed to complete their tasks within a 15-day period.

For annotating the misogyny memes, we consider four misogyny labels namely Stereotype, Objectification, Shaming, and Violence inspired by the work (Hakimov et al., 2022). Further, the annotators were asked to annotate the humor type, misogyny intensity, metaphor localization, metaphor object, and meme template. Each annotator independently labeled memes across the four categories. For classification tasks, we employed majority voting to determine final labels. We did a compensate the annotators at 1 BDT per sample. After completing the meme classification and detection annotation process, inspired from the work (Hwang and Shwartz, 2023), we also annotate the explanation of each meme. For completing the explanation annotation, we instructed annotators to write a short but comprehensive explanation for the memes, documenting structured metaphor analyses according to our standardized format. The explanation annotation guideline is detailed in Appendix Section B.1. For this task of meme explanation, we divided the 2,000 samples among annotators. Each meme was annotated by one annotator and compensating annotators at 2 BDT per sample. We provide the data validation process and metrics in Appendix D.

4 BANMIME : Dataset Statistics

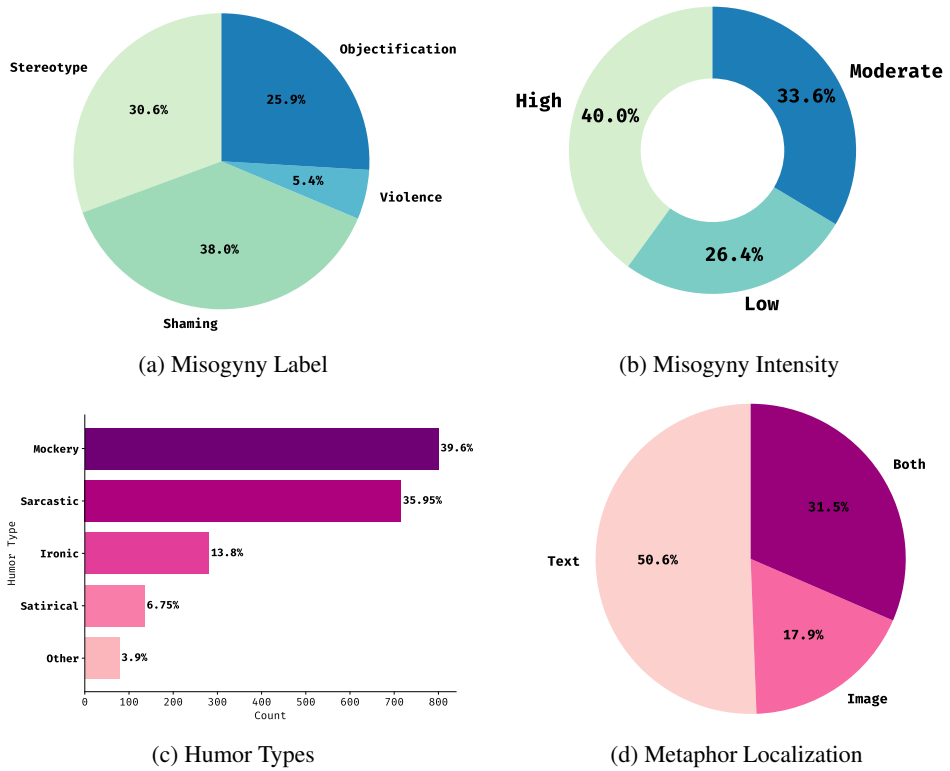
Meme Collection. The BANMIME dataset comprises 2,000 labeled Bangla memes collected from three major platforms: Facebook (1,300), Instagram (450), and Reddit (250). We implemented a stratified sampling approach to partition the dataset into training (1,500) and test (500) splits, preserving the distribution of categories across both sets. The corpus exhibits linguistic diversity, with text length ranging from 6 to 963 characters and word count varying from 3 to 150 words per meme. Table 1 presents the comprehensive statistics of the data set.

Source Distribution	**# Samples**
Facebook	1300
Instagram	450
Reddit	250
Splits	
- Train	1500
- Test	500
Text Statistics	
Mean Character Length	91.91
Max Character Length	963
Min Character Length	6
Mean Word Count	15.44
Max Word Count	150
Min Word Count	3

Table 1: Statistical overview of BANMIME dataset showing source distribution across platforms, training-testing splits, and text characteristics demonstrating linguistic diversity in Bangla memes.

Misogyny Label & Intensity. Analysis of misogynistic content distribution reveals distinct patterns specific to Bangla meme culture, as illustrated in Figure 3(a). Shaming and Stereotype constitute the majority of instances, followed by Objectification and Violence, which appears least frequently. This distribution suggests that while explicit violent content is comparatively uncommon in this medium, more subtle and culturally embedded forms of misogyny predominate. The intensity analysis depicted in Figure 3(b) demonstrates that high and moderate intensity content dominate the dataset.

Humor Categorization. Our analysis reveals distinct humor patterns across misogynistic content in Bangla memes, as visualized in Figure 3(c). Mockery and sarcasm constitute the vast majority of instances, while ironic, satirical, and other forms appear less frequently. The notable prevalence of mockery and sarcasm suggests that Bangla misogy-



(a) Misogyny Label

(b) Misogyny Intensity

(c) Humor Types

(d) Metaphor Localization

Figure 3: a) and (b) show the misogynistic content distribution in the BANMIME dataset. (c) and (d) show discourse analysis on humor and metaphore localization.

nistic memes predominantly employ direct ridicule rather than more complex humor forms, potentially amplifying their harmful impact.

Metaphor Analysis. Metaphorical expressions emerge as key mechanisms for conveying misogyny in Bangla memes. As shown in Figure 3(d), text-based metaphors predominate over multimodal and image-based varieties, highlighting the centrality of linguistic elements while also demonstrating sophisticated text-image integration. Common metaphor objects include bodies, character traits, sexual objectification, and domestic abilities.

Meme Explanation Analysis. To enhance qualitative understanding, we annotated comprehensive explanations for each meme. As shown in Table 2, these explanations vary in length and complexity, offering insights into cultural and linguistic mechanisms present in Bangla memes. These annotations document how specific cultural references, visual-textual interactions, and linguistic elements function within this multimodal format. This explanatory layer strengthens the dataset's utility for developing detection systems capable of understanding contextual nuances in multimodal content.

More data analysis on the meta data and meme template analysis for our BANMIME dataset can be found in Appendix E.

Statistics	Text
Mean Character Length	192.59
Max Character Length	725
Min Character Length	33
Mean Word Count	30.94
Max Word Count	120
Min Word Count	6

Table 2: Statistical analysis of meme explanations showing character and word count distributions that reveal cultural and linguistic mechanisms of misogynistic expression.

5 Experiment Design

To evaluate the understanding of visual metaphors and their role in detecting misogyny in memes on our BANMIME dataset, we select a diverse set of multimodal VLMs. We classify our experimentation designs into two categories: (i) Prompt-based Experiments and (ii) LoRA Finetuning Experiments.

5.1 Prompt-Based Experiments

For this experiments, we consider Zero Shot Prompting and Chain of Thought(CoT) prompting. In zero-shot setting the input is question text, the meme text, and multiple options, while the output is to predict the answer from the option set and explain metaphor. In CoT The model was encouraged to think step by step((e.g., identifying the metaphor object, explain the metaphor, classifying the misogyny category) before given the final answer. Refer to Appendix H for detailed prompt.

We have experimented prompt based approaches on pretrained VLMs LLaVA-1.5-7B (Liu et al., 2024a), Llama-3.2-11B-V (Grattafiori et al., 2024), LlaVA-NeXT-mistral-7b (Liu et al., 2024b), Phi-4-multimodal (Microsoft et al., 2025), Phi-3.5-V (Abdin et al., 2024), Gemma-3-12B (Team et al., 2025), and Qwen2.5-VL-7B- (Bai et al., 2025). In addition, we conduct evaluations using closed-source large VLMs, including GPT-4o (GPT-4o-mini), and Gemini (Gemini-2.0-Flash).

5.2 VLM LoRA Fine-Tuning

(Dehan et al., 2024) demonstrated that fine-tuning LLMs, particularly TinyLLMs, can exhibit strong performance in low-resource languages such as Bangla. Building on this, we also extend our experimentation to include LoRA fine-tuning for open-source models. Most open-source VLMs adopt a model architecture commonly referred to as the `vision_enc-adapter-LLM` pipeline. We apply LoRA (Hu et al., 2022) finetuning on the LLM part ϕ_{LLM} of the VLMs. We explore three distinct fine-tuning approaches, all operating on the same set of memes and instructions but differing in how it has been supervised to response.

Augmented Direct Prompt Fine-Tuning. In this approach, we enhance multiple-choice VQA training by permuting the position of the correct class label c^* within a fixed label set $\mathbf{C} = \{c_1, c_2, c_3, c_4\}$. For each input triplet $(\mathbf{I}, \mathbf{Q}, c^*)$, this results in four augmented variants, increasing data diversity while preserving the semantic meaning. Each sample, consisting of the instruction prompt $\mathbf{Inst}$, image $\mathbf{I}$, and class label set $\mathbf{C}$, is fed to the model.

Standard LoRA Fine-Tuning. In the standard setting, we input the instruction $\mathbf{Inst}$, question $\mathbf{Q}$, image $\mathbf{I}$, and the class label set $\mathbf{C}$ to the vision-language model. The training objective is to identify the correct class label $c^* \in \mathbf{C}$.

Chain-of-Thought Fine-Tuning. We also explore the CoT fine-tuning strategy to improve multimodal reasoning. In our context—detecting misogyny and identifying metaphors in memes—we adapt CoT fine-tuning to guide vision-language models through three key steps: (1) identifying metaphorical elements, (2) explaining their semantic roles, and (3) linking these interpretations to the final misogyny classification. Our dataset provides ground truth for both metaphor localization ($T_{\text{Localization}}$) and explanatory reasoning (T_{Exp}). Each training instance encourages the model to "think aloud" using guided prompts along with basic task instructions.

The model is trained to generate intermediate reasoning steps by identifying metaphorical clues ($T_{\text{metaClue}} \rightarrow T_{\text{Localization}}$), producing semantic explanations ($T_{\text{metaExp}} \rightarrow T_{\text{Exp}}$), and finally classifying the meme as c^*. The overall training objective is:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{class}} + \mathcal{L}_{\text{NTP}}^{\text{clue}} + \mathcal{L}_{\text{NTP}}^{\text{exp}},$$

where $\mathcal{L}_{\text{NTP}}$ represents next-token prediction losses. Due to space constraints, we provide details of the fine-tuning approaches in Appendix F.1, and outline our experimental setup and implementation specifics in Appendix F.2.

6 Result Analysis

The performance of different VLMs in classifying misogyny and generating explanations on our BANMIME dataset is reported in Table 3. Table 7 presents the VLMs' performance in detecting humor types within the dataset, while Table 8 details the performance in assessing misogyny intensity and metaphor localization. In Table 9 we also analyze the performance of the models across Meme Template.

Performance of Close Source Models For the closed-source VLMs, both Gemini and GPT show similar performance, with Gemini outperforming GPT by over 7% in the LAVE score under a zero-shot prompt. Interestingly, both models experience a performance decline when using the CoT prompt. Specifically, Gemini's performance drops by 3%, while GPT sees a larger decline of over 10%. Regarding explanation generation, Gemini's performance decreases by less than 1%, whereas GPT's drop exceeds 2%. The CoT prompting appears to negatively impact both models' performance in this regard. Notably, Gemini demonstrates stronger

Models	Misogyny Classification					Explanation		
	Sham	Stereo	Obj	Vio	Avg	BScore	LAVE	Avg
Zero Shot Prompt								
Closed Source VLMs								
Gemini2.0 Flash	34.97	52.55	65.77	72.41	**56.43**	86.70	**35.00**	**60.85**
GPT-4o-mini	26.26	61.33	60.33	**77.42**	56.34	**87.26**	27.20	57.23
Open Source VLMs								
Llama-3.2V 11B	12.31	18.12	42.50	16.13	22.27	82.82	0.35	41.59
Gemma-3-12B	52.02	22.67	48.76	67.74	47.80	83.54	8.00	45.77
Qwen2.5-VL 7B	22.73	41.33	57.02	35.48	39.14	78.60	0.60	39.60
Phi-3.5	23.23	67.33	25.62	6.45	30.66	82.30	0.23	41.27
Phi-4	49.49	28.86	20.83	16.13	28.83	81.63	0.20	40.92
LLaVA-1.5 7B	16.92	39.60	27.50	19.35	25.84	82.40	0.22	41.31
LLaVA-NeXT 7B	29.59	25.68	16.53	16.67	22.12	79.25	0.45	39.85
CoT Prompt								
Closed Source VLMs								
Gemini2.0 Flash	48.11	64.23	47.75	55.17	53.82	86.96	31.60	59.28
GPT-4o-mini	22.54	**80.69**	44.55	33.33	45.28	87.20	22.94	55.07
Open Source VLMs								
Llama-3.2V 11B	11.68	35.33	16.67	3.33	16.75	83.53	1.73	42.63
Gemma-3-12B	**75.25**	11.33	38.02	41.94	41.64	84.49	7.36	45.93
Qwen2.5-VL 7B	24.24	61.74	28.3 3	16.13	32.61	84.56	5.13	44.85
Phi-4	10.71	59.46	14.17	2.45	22.09	84.41	1.29	42.85
LLaVA-1.5 7B	7.61	57.72	10.08	6.45	20.47	84.34	0.27	42.31
LLaVA-NeXT 7B	8.67	43.62	11.57	6.45	17.58	84.23	2.16	43.20
LoRA (CoT) Fine-Tuning								
Llama-3.2V 11B	32.83	65.33	52.89	48.39	49.86	86.01	8.00	47.01
Gemma-3-12B	16.16	76.67	**49.59**	64.52	51.74	86.16	13.80	49.98
Qwen2.5-VL 7B	44.95	59.33	48.76	51.61	51.16	85.36	5.00	45.18
LLaVA-1.5 7B	24.62	48.33	52.98	54.26	45.05	85.19	4.80	44.20
Paligemma-2-10B	8.08	61.33	42.98	61.29	43.42	81.80	2.00	41.90

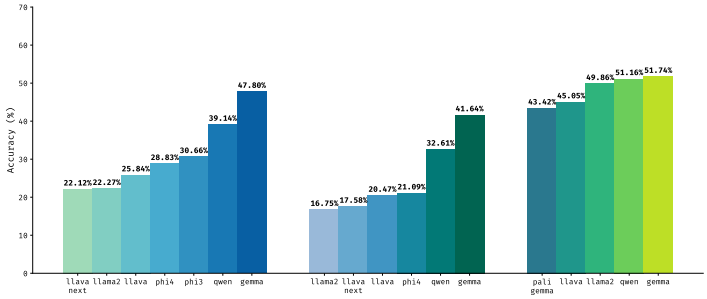
Table 3: The effect of different prompt techniques on the test split of the BANMIME dataset is reported in terms of accuracy percentage. Here, Sham, Stereo, Obj, and Vio stand for Shaming, Stereotype, Objectification, and Violence, respectively. Highest performance per category is colored in bold.

explanation capabilities, achieving the best LAVE score 60.85% with a zero-shot prompt. The model surpass the others in prompting experiments.

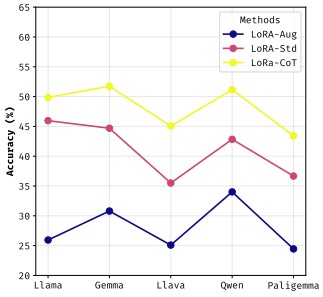
Performance of Open Source Models Among the open-source models, Gemma-3 outperforms the other VLMs. Qwen2.5 and Phi-3.5 show moderate performance, while the others perform the worst. Notably, except for Gemma-3, the LAVE scores of the other open VLMs are below 1, indicating their inability to provide meaningful reasoning. Interestingly, although the CoT prompting causes a decline in model performance, it leads to an improvement in the LAVE scores, particularly for Qwen2.5. However, Phi-4 experiences the most performance drop with CoT prompting. For the open VLMs, while zero-shot prompting yields better performance, the explanation quality remains poor. In contrast, although CoT prompting degrades performance, it enhances the quality of the generated explanations.

To What Extent Does Fine-Tuning Improve Performance? We assess the impact of LoRA-based fine-tuning by comparing model performance across Zero-Shot, CoT prompting, and fine-tuned settings. Overall, fine-tuning consistently improves classification accuracy across model families, including LLaVA, Phi, Qwen, Gemma, and PaLI-Gemma. Notably, open models still trail behind proprietary systems in absolute performance. A detailed comparison with accuracy trends and class-wise breakdowns is available in Appendix G.1.

Comparison on Different Finetuning Strategies As discussed in the section 5.2, we evaluate three settings: Standard fine-tuning $LoRA_{std}$, finetune with data augmentation $LoRA_{aug}$, and Chain-of-Thought supervised fine-tuning $LoRA_{CoT}$. In this section we answer the hypothesis we arise in section 5.2. The analysis reveals two central findings: (1) data augmentation does not consistently enhance performance, and (2) $LoRA_{CoT}$ tends to yield the most reliable gains across categories



(a) Model Performance Comparison. From left to right (Zero-Shot, CoT, LoRA-Finetune



(b) Finetune Result

Figure 4: Model performance comparison using different evaluation methods and the effect of finetuning.

Models	Misogyny categories				
	Shaming	Stereotype	Obj	Violence	Avg
Different Fine-Tuning Strategies					
LLaMa-3.2V 11B					
LoRA_{aug}	0.5	99.3	0.0	3.2	25.9
LoRA_{std}	11.6	61.3	49.6	61.3	46.0
LoRA_{CoT}	32.8	65.3	52.9	48.4	**49.9**
Gemma-3 12B					
LoRA_{aug}	8.1	98.0	7.4	9.7	30.8
LoRA_{std}	10.1	70.7	49.6	48.4	44.7
LoRA_{CoT}	16.2	76.7	49.6	64.5	**51.7**
Qwen2.5-VL 7B					
LoRA_{aug}	49.0	58.7	5.8	22.6	34.0
LoRA_{std}	47.0	59.0	27.3	37.1	42.6
LoRA_{CoT}	45.0	59.3	48.8	51.6	**51.2**
LLaVa-1.5 7B					
LoRA_{aug}	13.1	84.0	0.0	3.2	25.1
LoRA_{std}	8.1	46.7	38.8	48.4	35.5
LoRA_{CoT}	24.6	48.3	53.0	54.3	**45.1**
Paligemma-2 10B					
LoRA_{aug}	38.9	18.0	24.8	16.1	24.4
LoRA_{std}	9.6	39.3	39.7	58.1	36.7
LoRA_{CoT}	8.1	61.3	43.0	61.3	**43.4**

Table 4: Model performance on different finetune setups: LoRA-FT models are evaluated under different training setups. Here $LoRA_{CoT}$ refers to chain of though finetuning, $LoRA_{aug}$ = Finetune with augmentation, $LoRA_{std}$ = Standard finetuning with no augmentation.

and models. Detailed analysis is provided in Appendix G.2.

Metaphor Explanation Analysis We conduct a qualitative analysis to assess how well our fine-tuned model interprets metaphors in misogyny categorization tasks. The model shows promising reasoning abilities, often aligning with human annotators, particularly due to CoT fine-tuning. Notably, errors are frequently linked to misinterpretation of metaphorical content. Detailed examples and additional discussion are provided in Appendix G.3.

Model Performance Across Misogyny Categories We observe highly skewed results across the misogyny categories. Most open VLMs struggle to accurately predict labels for Shaming and Violence memes. In the Shaming category, all models, except Gemma-3 and Phi-4, perform below 35% in zero-shot settings. However, CoT prompting helps Gemma-3 and Gemini improve their performance (a 33% improvement for Gemma-3 and 13% for Gemini), while the performance of the other models deteriorates. Fine-tuning improves the performance of LLaMa, Qwen, and LLaVa in this category. For Stereotype memes, GPT with CoT achieves over 80% accuracy, while Gemma-3 with CoT performs the worst. Overall, most models perform better at detecting stereotype memes than at detecting other categories. In the Objectification category, Gemma-3 with CoT-FT delivers the best result, achieving nearly 70% accuracy. GPT with zero-shot performs well at detecting violence-related memes, but the introduction of CoT leads to a large performance drop across all models. Most open-source VLMs, particularly Gemma-3, struggle to detect violence memes. However, for most models (except Gemma-3), fine-tuning results in improved performance.

Model Performance on Humor Category Table 7 presents a comparative analysis of VLMs in detecting five types of humor—Ironic, Mockery, Satirical, Sarcastic, and Other. Closed-source models like Gemini2.0 Flash and GPT-4o-mini typically outperform open-source models in humor detection, with Gemini2.0 Flash leading the way in both Zero Shot and CoT prompting. Among open-source models, Gemma-3-12B stands out, particularly excelling in Mockery and Satirical humor. The Chain of Thought method enhances Gemini2.0 Flash's performance, while GPT-4o-mini sees a slight decline. Fine-tuning with LoRA boosts the performance of Gemma-3-12B and Qwen2.5-VL 7B, especially in detecting Mockery and Sarcastic humor. However,

models like LLaVA-1.5 7B and LLaVA-NeXT 7B consistently perform poorly across all humor categories, showing limited improvement even with fine-tuning. In general, Gemini2.0 Flash excels, particularly in Ironic and Sarcastic humor, making it the top performer across all categories. For further details, refer to Appendix Section G.4.

Impact of Misogyny Intensity Table 8 summarizes VLM performance on Misogyny Intensity detection across Zero Shot, CoT, and LoRA (CoT) Fine-Tuning. Gemini2.0 Flash consistently leads, scoring 49.41 in Zero Shot and 53.66 with CoT. GPT-4o-mini follows closely. Among open-source models, Gemma-3-12B performs best, improving from 33.78 (Zero Shot) to 45.24 (LoRA). Qwen2.5-VL 7B also shows gains, while models like Llama-3.2V and LLaVA variants lag behind.

Effect of Metaphor Localization In terms of Metaphor Localization, Gemini2.0 Flash again proves to be the leader across all prompting methods, with an average of 50.07 in Zero Shot and 54.11 in CoT. Gemma-3-12B is strong in Zero Shot (avg: 37.57) and shows further improvement under CoT (avg: 40.46), although it still trails behind Gemini2.0 Flash. LoRA Fine-Tuning brings noticeable improvements for Gemma-3-12B and Qwen2.5-VL 7B, with average scores of 49.36 and 38.83, respectively. However, Llama-3.2V 11B still struggles in this task, with a low average of 26.91. Overall, Gemini2.0 Flash remains the top performer across both tasks, while Gemma-3-12B benefits most from fine-tuning. For a more detailed explanation, can be found Appendix Section G.5.

Influence of Meme Template. The Table 9 compares the performance of VLMs across meme templates. Gemini2.0 Flash consistently outperforms all models across prompting methods, particularly excelling in Doge and Troll Face, achieving high average scores, especially in Chain of Thought prompting. GPT-4o-mini also shows strong performance, particularly in Wojak and Doge, with high averages across all settings. Open-source models such as Llama-3.2V 11B and Gemma-3-12B perform reasonably well, with Llama-3.2V 11B showing the strongest results among them, especially for Doge and Troll Face. Fine-tuning via LoRA enhances performance, with Llama-3.2V 11B and Paligemma-2-10B achieving notable improvements in categories like Doge. Overall, closed-source models outperform open-source models, but fine-tuning helps close the gap for several open-source VLMs. More details can be found in Appendix Section G.6.

7 Conclusion

We present BANMIME , the first culturally grounded dataset for misogyny detection and explanation in Bangla memes, enriched with fine-grained annotations and detailed metaphor-based reasoning. Our extensive experiments reveal that current VLMs including advanced closed-source VLMs like GPT and Gemini—struggle to interpret complex multimodal cues, especially those involving metaphor. These models often fail to generate coherent explanations and fall significantly short of accurate prediction. We envision our work will drive the development of culturally-aware multimodal systems for Bangla content moderation and foster improved reasoning abilities in vision-language models for underrepresented languages.

Limitations

A primary limitation of our study is the relatively small dataset size (2,000 samples), constrained by the need for high-quality, culturally informed human annotations—particularly for nuanced tasks like metaphor explanation and misogyny categorization. Given the complexity of code-mixed Bangla-English memes and the visual-textual reasoning required, annotation was both time-consuming and resource-intensive. While our fine-tuned models show strong performance, the limited scale may impact generalizability. Future work should explore scalable annotation strategies, such as active learning or semi-automated pipelines, to expand the dataset without compromising quality.

Ethical Statement

We excluded any memes containing explicit nudity to maintain ethical content standards. No personal information was collected from annotators, and no personally identifiable information (PII) is included in the dataset. We conducted a thorough review of the data to identify and mitigate potential biases. While our models are designed with fairness in mind, we acknowledge that biases—particularly in subjective tasks like abusive content detection—can be difficult to eliminate entirely. Labeling such content can be inherently

subjective, and although we aimed for objectivity, some unintended biases may exist. Nonetheless, our dataset exhibits strong annotation quality, supported by high inter-annotator agreement. All datasets and models will be publicly released (where permissible) to ensure transparency and reproducibility, with full credit to original sources. We emphasize that any biases in our data are unintentional, and we have no intent to harm any individual or group. Further evaluation may be necessary to assess and address any residual model bias in downstream applications.

Authors' Contribution

Ayon and Sayma initiated the project with the goal of creating a dataset for detecting misogyny in memes and reported their progress to Fahim. Fahim proposed incorporating explanations and analyzing SOTA VLMs for both meme detection and explanation generation. Ayon and Sayma developed the annotation guidelines for explanation generation and provided guidance to the annotators. Following the creation of the dataset, Fahim directed Moshiur to carry out the necessary experiments and also contributed to several visualizations in the paper. Moshiur and Fuad executed all the experiments and reported their findings to Fahim. Fahim led the project, co-wrote and updated the paper alongside Ayon and Moshiur. Ibrahim and Mahabubur supervised the project and provided funding support.

Acknowledgments

We are thankful to Independent University, Bangladesh, for their support of this project. We would also like to express our gratitude to the Center for Computational & Data Sciences (CCDS Lab) for providing computational facilities and supervising this project.

References

Marah Abdin, Jyoti Aneja, Hany Awadalla, Ahmed Awadallah, Ammar Ahmad Awan, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Jianmin Bao, Harkirat Behl, Alon Benhaim, Misha Bilenko, Johan Bjorck, Sébastien Bubeck, ..., Yi Zhang, Yue Zhang, Yunan Zhang, and Xiren Zhou. 2024. Phi-3 technical report: A highly capable language model locally on your phone. *Preprint*, arXiv:2404.14219.

Shawly Ahsan, Eftekhar Hossain, Omar Sharif, Avishek Das, Mohammed Moshiul Hoque, and M. Dewan. 2024. A multimodal framework to detect target aware aggression in memes. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2487–2500, St. Julian's, Malta. Association for Computational Linguistics.

Maria Anzovino, Elisabetta Fersini, and Paolo Rosso. 2018. Automatic identification and classification of misogynistic language on twitter. In *Natural Language Processing and Information Systems: 23rd International Conference on Applications of Natural Language to Information Systems, NLDB 2018, Paris, France, June 13-15, 2018, Proceedings 23*, pages 57–64. Springer.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. 2025. Qwen2.5-vl technical report. *Preprint*, arXiv:2502.13923.

Khandis R Blake, Siobhan M O'Dean, James Lian, and Thomas F Denson. 2021. Misogynistic tweets correlate with violence against women. *Psychological science*, 32(3):315–325.

Farhan Noor Dehan, Md Fahim, AKM Mahabubur Rahman, M Ashraful Amin, and Amin Ahsan Ali. 2024. Tinyllm efficacy in low-resource language: An experiment on bangla text classification task. In *International Conference on Pattern Recognition*, pages 472–487. Springer.

Md Fahim. 2023. Aambela at blp-2023 task 2: Enhancing banglabert performance for bangla sentiment analysis task with in task pretraining and adversarial weight perturbation. In *Proceedings of the First Workshop on Bangla Language Processing (BLP-2023)*, pages 317–323.

Md Fahim, Fariha Shifat, Fabiha Haider, Deeparghya Barua, Md Sourove, Md Ishmam, and Md Bhuiyan. 2024. Banglatlit: A benchmark dataset for back-transliteration of romanized bangla. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 14656–14672.

Elisabetta Fersini, Francesca Gasparini, Giulia Rizzi, Aurora Saibene, Berta Chulvi, Paolo Rosso, Alyssa Lees, and Jeffrey Sorensen. 2022. Semeval-2022 task 5: Multimedia automatic misogyny identification. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, pages 533–549.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, ... Zhenyu Yang, Zhiwei Zhao, and Zhiyu Ma. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Qin Gu, Nino Meisinger, and Anna-Katharina Dick. 2022. QiNiAn at SemEval-2022 task 5: Multi-modal misogyny detection and classification. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, pages 736–741, Seattle, United States. Association for Computational Linguistics.

Fabiha Haider, Fariha Tanjim Shifat, Md Farhan Ishmam, Deeparghya Dutta Barua, Md Sakib Ul Rahman Sourove, Md Fahim, and Md Farhad Alam. 2024. Banth: A multi-label hate speech detection dataset for transliterated bangla. *arXiv preprint arXiv:2410.13281*.

Sherzod Hakimov, Gullal Singh Cheema, and Ralph Ewerth. 2022. TIB-VA at SemEval-2022 task 5: A multimodal architecture for the detection and classification of misogynous memes. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, pages 756–760, Seattle, United States. Association for Computational Linguistics.

Eftekhar Hossain, Omar Sharif, and Mohammed Moshiul Hoque. 2022a. Memosen: A multimodal dataset for sentiment analysis of memes. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 1542–1554.

Eftekhar Hossain, Omar Sharif, and Mohammed Moshiul Hoque. 2022b. Mute: A multimodal dataset for detecting hateful memes. In *Proceedings of the 2nd conference of the asia-pacific chapter of the association for computational linguistics and the 12th international joint conference on natural language processing: student research workshop*, pages 32–39.

Eftekhar Hossain, Omar Sharif, and Mohammed Moshiul Hoque. 2022c. MUTE: A multimodal dataset for detecting hateful memes. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing: Student Research Workshop*, pages 32–39, Online. Association for Computational Linguistics.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. 2022. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3.

EunJeong Hwang and Vered Shwartz. 2023. MemeCap: A dataset for captioning and interpreting memes. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 1433–1445, Singapore. Association for Computational Linguistics.

Alvi Md Ishmam and Sadia Sharmin. 2019. Hateful speech detection in public facebook pages for the bengali language. In *2019 18th IEEE International Conference On Machine Learning And Applications (ICMLA)*, pages 555–560.

Khondoker Ittehadul Islam, Sudipta Kar, Md Saiful Islam, and Mohammad Ruhul Amin. 2021. Sentnob: A dataset for analysing sentiment on noisy bangla texts. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3265–3271.

Sarif Sultan Saruar Jahan, Raqeebir Rab, Peom Dutta, Hossain Muhammad Mahdi Hassan Khan, Muhammad Shahariar Karim Badhon, Sumaiya Binte Hassan, and Ashikur Rahman. 2023. Deep learning based misogynistic bangla text identification from social media. *Computing and Informatics*, 42(4):993–1012.

Douwe Kiela, Hamed Firooz, Aravind Mohan, Vedanuj Goswami, Amanpreet Singh, Pratik Ringshia, and Davide Testuggine. 2020. The hateful memes challenge: Detecting hate speech in multimodal memes. *Advances in neural information processing systems*, 33:2611–2624.

Gokul Karthik Kumar and Karthik Nandakumar. 2022. Hate-CLIPper: Multimodal hateful meme classification based on cross-modal interaction of CLIP features. In *Proceedings of the Second Workshop on NLP for Positive Impact (NLP4PI)*, pages 171–183, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the 29th Symposium on Operating Systems Principles*, pages 611–626.

Hongzhan Lin, Ziyang Luo, Wei Gao, Jing Ma, Bo Wang, and Ruichao Yang. 2024. Towards explainable harmful meme detection through multimodal debate between large language models. In *Proceedings of the ACM Web Conference 2024*, pages 2359–2370.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2024a. Improved baselines with visual instruction tuning. *Preprint*, arXiv:2310.03744.

Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. 2024b. Llava-next: Improved reasoning, ocr, and world knowledge.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. *Preprint*, arXiv:2304.08485.

Pan Lu, Swaroop Mishra, Tony Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. 2022. Learn to explain: Multimodal reasoning via thought chains for science question answering. *Preprint*, arXiv:2209.09513.

Sidharth Mahesh, Sonith D, Gauthamraj Gauthamraj, Kavya G, Asha Hegde, and H Shashirekha. 2024. MUCS@LT-EDI-2024: Exploring joint representation for memes classification. In *Proceedings of the*

Fourth Workshop on Language Technology for Equality, Diversity, Inclusion, pages 282–287, St. Julian's, Malta. Association for Computational Linguistics.

Oscar Mañas, Benno Krojer, and Aishwarya Agrawal. 2024. Improving automatic vqa evaluation using large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 4171–4179.

Microsoft, :, Abdelrahman Abouelenin, Atabak Ashfaq, Adam Atkinson, Hany Awadalla, Nguyen Bach, Jianmin Bao, Alon Benhaim, Martin Cai, Vishrav Chaudhary, Congcong Chen, Dong Chen, Dongdong Chen, Junkun Chen, Weizhu Chen, Yen-Chun Chen, Yi ling Chen, Qi Dai, Xiyang Dai, Ruchao Fan, Mei Gao, Min Gao, Amit Garg, Abhishek Goswami, Junheng Hao, Amr Hendy, Yuxuan Hu, Xin Jin, Mahmoud Khademi, Dongwoo Kim, Young Jin Kim, Gina Lee, Jinyu Li, Yunsheng Li, Chen Liang, Xihui Lin, Zeqi Lin, Mengchen Liu, Yang Liu, Gilsinia Lopez, Chong Luo, Piyush Madan, Vadim Mazalov, Arindam Mitra, Ali Mousavi, Anh Nguyen, Jing Pan, Daniel Perez-Becker, Jacob Platin, Thomas Portet, Kai Qiu, Bo Ren, Liliang Ren, Sambuddha Roy, Ning Shang, Yelong Shen, Saksham Singhal, Subhojit Som, Xia Song, Tetyana Sych, Praneetha Vaddamanu, Shuohang Wang, Yiming Wang, Zhenghao Wang, Haibin Wu, Haoran Xu, Weijian Xu, Yifan Yang, Ziyi Yang, Donghan Yu, Ishmam Zabir, Jianwen Zhang, Li Lyna Zhang, Yunan Zhang, and Xiren Zhou. 2025. Phi-4-mini technical report: Compact yet powerful multimodal language models via mixture-of-loras. *Preprint*, arXiv:2503.01743.

Abrar Shadman Mohammad Nahin, Isfara Islam Roza, Tasnuva Tamanna Nishat, Afia Sumya, Hanif Bhuiyan, and Md Moinul Hoque. 2024. 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)*, pages 01–06. IEEE.

Marinella Paciello, Francesca D'Errico, Giorgia Saleri, and Ernestina Lamponi. 2021. Online sexist meme and its effects on moral and emotional processes in social media. *Computers in human behavior*, 116:106655.

Mohammad Zia Ur Rehman, Sufyaan Zahoor, Areeb Manzoor, Musharaf Maqbool, and Nagendra Kumar. 2025. A context-aware attention and graph neural network-based multimodal framework for misogyny detection. *Information Processing & Management*, 62(1):103895.

Saima Sadiq, Arif Mehmood, Saleem Ullah, Maqsood Ahmad, Gyu Sang Choi, and Byung-Won On. 2021. Aggression detection through deep neural model on twitter. *Future Generation Computer Systems*, 114:120–129.

Aakash Singh, Deepawali Sharma, and Vivek Kumar Singh. 2024. Mimic: misogyny identification in multimodal internet content in hindi-english code-mixed language. *ACM Transactions on Asian and Low-Resource Language Information Processing*.

Smriti Singh, Amritha Haridasan, and Raymond Mooney. 2023. "female astronaut: Because sandwiches won't make themselves up there": Towards multimodal misogyny detection in memes. In *The 7th Workshop on Online Abuse and Harms (WOAH)*, pages 150–159.

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

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, ... Alek Andreev, Cassidy Hardin, Robert Dadashi, and Léonard Hussenot. 2025. Gemma 3 technical report. *Preprint*, arXiv:2503.19786.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2023. Chain-of-thought prompting elicits reasoning in large language models. *Preprint*, arXiv:2201.11903.

Fan Wu, Bin Gao, Xiaoou Pan, Linlin Li, Yujiao Ma, Shutian Liu, and Zhengjun Liu. 2024. Fuser: an enhanced multimodal fusion framework with congruent reinforced perceptron for hateful memes detection. *Information Processing & Management*, 61(4):103772.

Savvas Zannettou, Tristan Caulfield, Jeremy Blackburn, Emiliano De Cristofaro, Michael Sirivianos, Gianluca Stringhini, and Guillermo Suarez-Tangil. 2018. On the origins of memes by means of fringe web communities. In *Proceedings of the internet measurement conference 2018*, pages 188–202.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

Yaowei Zheng, Richong Zhang, Junhao Zhang, Yanhan Ye, Zheyan Luo, Zhangchi Feng, and Yongqiang Ma. 2024. Llamafactory: Unified efficient fine-tuning of 100+ language models. *arXiv preprint arXiv:2403.13372*.

A Annotation Guidelines

This section provides comprehensive guidelines developed for annotating BANMIME dataset. Annotators were instructed to examine both visual and textual elements in each meme to identify instances of stereotype, objectification, shaming, and

(a) Stereotype

(b) Objectification

(c) Shaming

(d) Violence

Figure 5: Examples of the four types of misogynistic content in the BANMIME dataset, showing representative instances of Stereotype, Objectification, Shaming, and Violence with their corresponding metadata annotations.

violence against women. Figure 5 illustrates examples of memes representing each of these four categories, along with their complete metadata annotations including misogyny label, humor type, metaphor localization, textual content, meme explanation, and misogynistic intensity. The design of annotation guidelines is inspired by BanglaTLit paper (Fahim et al., 2024).

A.1 General Instructions

The annotation process followed these core principles:

- **Multimodal Analysis:** Annotators analyzed both image and text components, considering their semantic interrelationship and cultural context.
- **Mutually Exclusive Classification:** Each meme was assigned exactly one of the four primary categories based on its predominant misogynistic characteristics.
- **Comprehensive Metadata:** Annotators provided all required metadata alongside the primary label for multidimensional analysis.
- **Annotation Consistency:** Definitions were applied uniformly across the dataset to ensure reliability and reproducibility.
- **Cultural Relevance:** Bangla-specific cultural nuances and expressions were carefully considered during interpretation.

A.2 Primary Categories and Definitions

We developed a taxonomy comprising four distinct categories of misogynistic content adapted to reflect Bangla cultural contexts.

- **Stereotype**
  Definition: Content that attributes generalized and oversimplified characteristics to women, reinforcing traditional gender roles and limiting perceptions of women's abilities and behaviors.
  Indicators:
  - Portraying women as primarily interested in shopping/makeup
  - Depicting women as emotional, irrational, or intellectually inferior
  - Suggesting women are naturally suited only for domestic roles

- Attributing universal traits to all women

- **Objectification**
  Definition: Content that treats women as objects devoid of agency, focusing solely on their physical appearance or sexual attributes, thereby reducing their personhood.
  Indicators:
  - Reducing women to body parts
  - Comparing women to consumable items
  - Depicting women as decoration or aesthetic objects
  - Treating women as interchangeable or collectible

- **Shaming**
  Definition: Content that criticizes or mocks women for behaviors or characteristics that deviate from societal norms, often targeting aspects like sexuality, appearance, or lifestyle choices.
  Indicators:
  - Mocking women's appearance
  - Criticizing women's clothing/fashion choices
  - Ridiculing women's career or education choices
  - Shaming women for perceived promiscuity or prudishness

- **Violence**
  Definition: Content that promotes, normalizes, or makes light of physical, sexual, or psychological harm directed at women, including threats, coercion, or dehumanizing language.
  Indicators:
  - Jokes about domestic abuse
  - Normalizing sexual harassment or assault
  - Threatening language toward women
  - Dehumanizing depictions of women

B Metadata Annotation Schema

To enable multidimensional analysis of misogynistic content, annotators provided the following structured metadata for each meme:

- **Misogynistic Intensity**
  A three-point scale measuring severity:
  - **High:** Explicit, severe, and unambiguous misogyny
  - **Moderate:** Clear but less extreme misogynistic content
  - **Low:** Subtle or implicit misogynistic elements

- **Humor Type**
  Categorization of rhetorical mechanisms:
  - **Mockery:** Ridiculing or belittling women misogyny
  - **Sarcastic:** Using irony to convey contempt
  - **Ironic:** Expressing meaning opposite to the literal meaning
  - **Satirical:** Using humor to criticize aspects of society
  - **other:** Humor types not covered by the categories above

- **Metaphor Localization**
  Location of metaphorical content:
  - **Text:** Metaphor present in the textual content
  - **Image:** Metaphor conveyed through visual elements
  - **Both:** Metaphor is expressed in both text and imagery

- **Metaphor Object**
  Our analysis identified numerous metaphorical targets, with varying frequencies across the dataset. The most common metaphor objects include women's body, marriage/relationship, women's character, sexual objectification, women's intelligence, presidency, makeup, money, pregnancy, cooking ability, and sexual performance, among others.

- **Meme Template**
  Classification of formal structure:
  - **Non-Templated Memes:** Original meme without standard templates
  - **Established Templates:** Including Troll Face, Wojak, Rage Comic, Soyjak, Doge, and others

B.1 Meme Explanation Annotation

For the meme explanation annotation phase, annotators were instructed to write the explanation of each meme using a standardized format: Metaphor object + Metaphor location + Explanation. For example, in a meme containing luxury car references, the annotation might read: "Metaphor: Text, Metaphor Object: BMW, Meme Explanation: In the text, BMW, a popular and expensive car, mentioned in the text of the meme metaphorically indicates that a beautiful woman is a financial trap." This structured documentation approach enabled systematic analysis of how metaphors function within misogynistic discourse in Bangla memes, revealing patterns in the symbolic associations used to reinforce stereotypes, objectification, shaming, and violence against women.

C Annotation Tool

We developed a custom web-based annotation platform to facilitate systematic and consistent annotation of misogynistic memes in Bangla. Figure 7 illustrates the user interface, which incorporates several key components designed for efficient multi-class annotation:

- **Authentication and Configuration:** The left sidebar provides secure authentication, dataset parameter configuration, and annotation initialization. Access controls ensure that only authorized annotators can submit classifications.
- **Image Display Module:** The central panel presents memes at their original resolution, allowing annotators to examine both visual and textual elements in their native context.
- **Classification Interface:** A structured dropdown menu enforces the mutually exclusive nature of our annotation scheme, requiring annotators to select exactly one of the four predefined categories.
- **Contextual Documentation:** A dedicated text area allows annotators to document relevant observations and metaphor explanations according to the standardized format described in the guidelines.
- **Progress Management:** Intuitive navigation controls facilitate seamless movement between samples, with persistent progress tracking and automatic state saving.
- **Structured Data Export:** The platform supports JSON export for annotations, enabling seamless integration with validation processes and analytical pipelines.

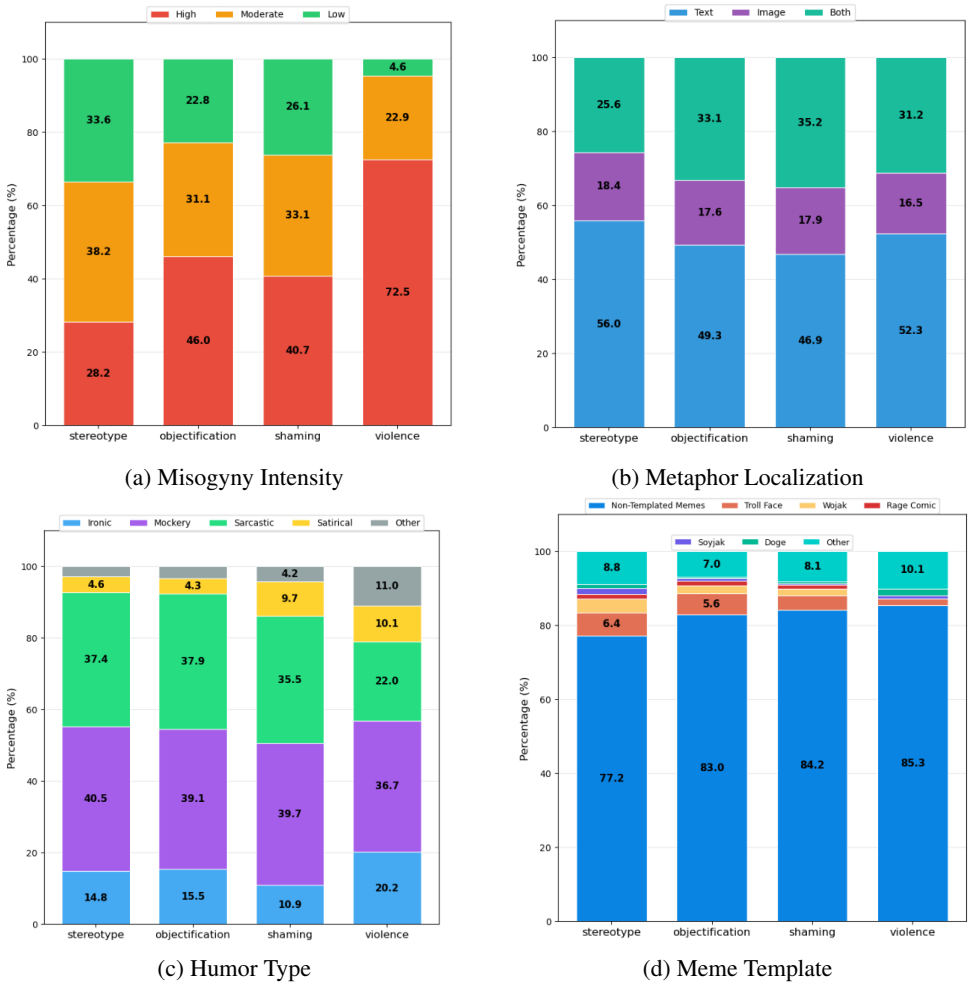
D Data Validation

Data Validation for Meme Identification. Inter-annotator reliability for the primary classification task was assessed using Cohen's kappa coefficients, as presented in Table 5. The results demonstrate substantial agreement across all misogyny categories. Notably, while prior research in content moderation reported κ scores of approximately 0.53, indicating moderate agreement (Islam et al., 2021), our refined annotation guidelines and annotator selection criteria yielded higher agreement scores. These robust agreement metrics across multiple evaluation methods affirm the clarity of our taxonomy and the effectiveness of our annotation protocol. The rigorous annotation methodology ensures that the BANMIME dataset establishes a reliable foundation for computational analysis of misogynistic discourse in Bangla memes.

Category	**Kappa(κ)**
Stereotype	0.76
Objectification	0.69
Shaming	0.71
Violence	0.78
Average	**0.74**

Table 5: Inter-annotator agreement for misogyny categories measured by Cohen's Kappa, with average score of 0.74 demonstrating substantial cross-category annotation reliability.

Data Validation for Meme Explanation. To evaluate the quality and consistency of meme explanations, we randomly selected a subset of 200 samples for independent analysis by each annotator. The similarity between these annotator-generated explanations was assessed using multiple automated metrics. The explanations exhibited exceptional consistency, with high ROUGE scores (ROUGE-1(F1): 86.72%, ROUGE-2(F1): 52.14%, ROUGE-L(F1): 85.96%) demonstrating substantial agreement in both content structure and coverage. Complementary evaluation using BLEU (74.33%), BERT similarity (95.87%), and METEOR (82.51%) metrics further confirmed strong semantic coherence and lexical alignment across the annotators' interpretations, validating the reliability of our explanation annotation approach.



(a) Misogyny Intensity

(b) Metaphor Localization

(c) Humor Type

(d) Meme Template

Figure 6: Cross-categorical analysis of metadata distributions in the BANMIME dataset: (a) Misogyny Intensity showing prevalence of high intensity content in violence categories; (b) Metaphor Localization patterns revealing predominance of text-based metaphors across all categories; (c) Humor Type distribution highlighting mockery as the primary vehicle for misogynistic content; and (d) Meme Template analysis demonstrating the overwhelming predominance of non-templated, locally-generated content across all misogyny types.

E Metadata Distribution Analysis Across Misogyny Categories

Meme Template Analysis.

Statistics	#Samples
Non-Templated Memes	1636
Troll Face	99
Wojak	47
Rage Comic	23
Soyjak	17
Doge	15
Other	163

Table 6: Meme template analysis shows non-templated content (81.8%) substantially outnumbers established templates (18.2%), indicating locally-generated content prevalence in Bangla misogynistic memes.

The distribution of meme templates presented in Table 6 reveals a notable pattern: non-templated memes substantially outnumber established templates. Among the established templates, Troll Face appears most frequently, followed by Wojak, Rage Comic, Soyjak, and Doge, with various other templates completing the distribution. This predominance of non-templated content represents a departure from global meme culture, suggesting that Bangla misogynistic meme creation is less influenced by international formats and more deeply rooted in locally generated content.

Meta Data Statistics To better understand how linguistic and visual elements interact within different types of misogynistic content, we present cross-categorical analyses of key metadata dimensions from the BANMIME dataset. These visualizations reveal distinctive patterns in how misogyny

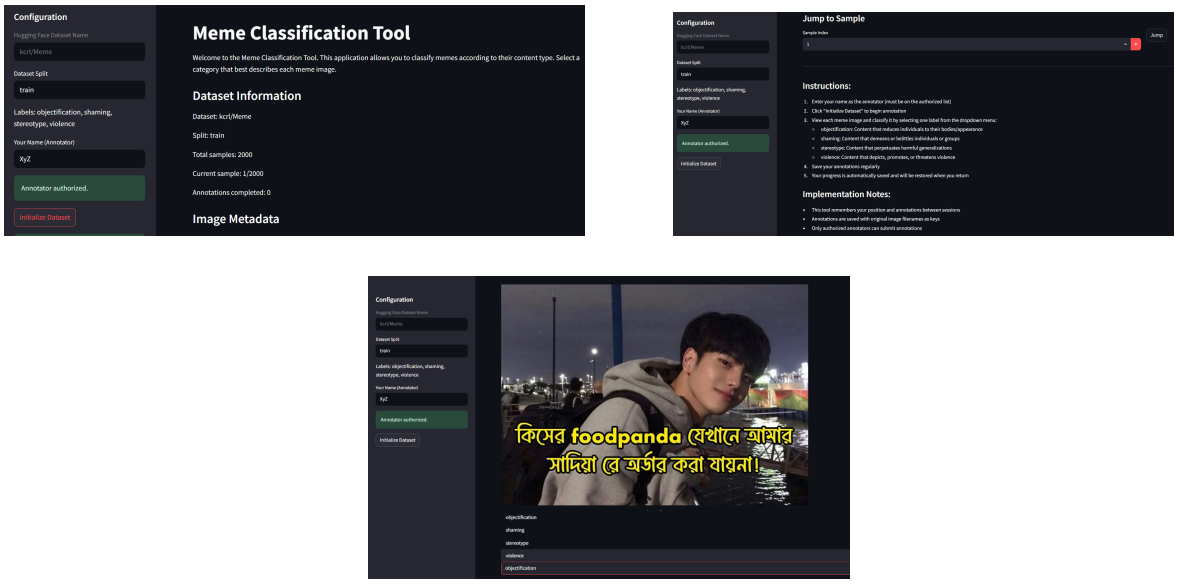


Figure 7: Interface of the web-based annotation tool developed for the BANMIME dataset, showing the configuration interface, annotation workspace, and meme display with classification options.

manifests across stereotype, objectification, shaming, and violence categories. Figure 6(a) presents the distribution of misogyny intensity levels (high, moderate, low) within each misogyny category. Violence exhibits the highest proportion of high-intensity instances (72.5%), while stereotype content shows a more even distribution with greater representation of low intensity (33.6%). This visualization quantifies the varying severity patterns characteristic of each misogyny type. Figure 6(b) illustrates metaphor localization patterns (text, image, both) across the four misogyny categories. Text-based metaphors predominate across all categories (46.9%-56.0%), though with notable variations. Stereotype content shows the highest proportion of combined text-image metaphors (25.6%), highlighting how stereotypical content often requires multimodal interpretation. Figure 6(c) visualizes humor type distribution (ironic, mockery, sarcastic, satirical, other) across categories, revealing that violence-themed content employs more ironic humor (20.2%) compared to other categories, while mockery remains the consistent primary vehicle across stereotype (40.5%) and objectification (39.1%) categories. Figure 6(d) demonstrates meme template usage patterns across misogyny categories, with the consistency of non-templated content (77.2%-85.3%) revealing a notable departure from global meme culture and suggesting greater reliance on locally generated content across all forms of misogynistic expression. These cross-categorical metadata distributions provide essential insights for developing more nuanced detection mechanisms capable of identifying the specific patterns associated with different manifestations of misogyny in Bangla memes.

F Experiment Design Details

F.1 Finetuning Approaches

Augmentation Strategy: Consider a data sample represented as (I, Q, M), where $\mathbf{I}$ is the image, $\mathbf{Q}$ is the associated question, and $\mathbf{M}$ is the correct answer selected from a set of class labels.

In our approach, we provide the Vision-Language Models (VLMs) with the question $\mathbf{Q}$, image $\mathbf{I}$, and the set of class labels $\mathbf{C} = \{c_1, c_2, c_3, c_4\}$, where c_1, c_2, c_3, c_4 represent the four possible options for the given question. The correct label c^* is one of the elements in $\mathbf{C}$.

To enhance the dataset and mitigate the scarcity of multiple-choice data, we perform augmentation by shuffling the position of the correct label c^* within the class label set $\mathbf{C}$. This generates four distinct permutations of the class label set, resulting in four new question-answer pairs for each original sample. Each augmented pair corresponds to a different arrangement of the class labels with the correct answer c^* placed in various positions.

For instance, consider a sample (I, Q, M) where the original class label set is $\mathbf{C} = \{c_1, c_2, c_3, c_4\}$ and the correct label is $c^* = c_1$. We have four labels for misogyny identifications. The augmented question-answer pairs generated from this are:

1. **Sample 1:** Question Q with class labels $\mathbf{C} = \{c_1, c_2, c_3, c_4\}$, answer: c_1
2. **Sample 2:** Question Q with class labels $\mathbf{C} = \{c_2, c_1, c_3, c_4\}$, answer: c_1
3. **Sample 3:** Question Q with class labels $\mathbf{C} = \{c_3, c_2, c_1, c_4\}$, answer: c_1
4. **Sample 4:** Question Q with class labels $\mathbf{C} = \{c_4, c_2, c_3, c_1\}$, answer: c_1

This process is repeated for every question-answer pair in the dataset. Each question-answer pair generates four distinct versions after augmentation, as the correct label c^* is shuffled into different positions within the class label set $\mathbf{C}$. If the dataset initially contains N samples, after applying this augmentation strategy, the total number of augmented samples will be $4N$, where N is the number of original samples and 4 represents the four distinct permutations per sample.

Chain-of-Thought (CoT) Fine-Tuning for Structured Reasoning: Chain-of-thought (CoT) reasoning refers to a structured paradigm wherein models generate intermediate steps or rationales that bridge the gap between input stimuli and the final prediction (Wei et al., 2023). In the context of multimodal question answering, previous work has demonstrated that including CoT demonstrations—either in the form of pre-answer rationales or post-answer explanations—can significantly improve the model's ability to decompose complex reasoning tasks into simpler, interpretable subproblems (Liu et al., 2023) (Lu et al., 2022).

We extend this paradigm to our multimodal meme understanding task, which involves detecting misogynistic content and identifying metaphorical usage. Each data sample is defined as $(\mathbf{I}, \mathbf{Q}, \mathbf{C}, c^*)$, where $\mathbf{I}$ is the meme image, $\mathbf{Q}$ is a question about the meme, $\mathbf{C} = \{c_1, c_2, c_3, c_4\}$ is the set of candidate labels, and $c^* \in \mathbf{C}$ is the correct class label.

In addition to the classification objective, our dataset includes ground-truth metaphor localization $T_{\text{Localization}}$ and explanation annotations T_{Exp}, which provide richer supervision signals. To this end, we adapt CoT fine-tuning to guide the model through a structured reasoning process comprising:

1. Identifying metaphorical clues in the image-text meme T_{metaClue},
2. Generating a textual explanation of their meaning T_{metaExp},
3. Predicting the correct misogyny category $c^* \in \mathbf{C}$.

Each training instance is framed as a sequence generation task where the model is prompted to "think aloud." Given an instruction prompt $\mathbf{Inst}$, the model generates:

$$\begin{aligned}\text{Model Output: } & T_{\text{metaClue}} \rightarrow T_{\text{Localization}}, \\ & T_{\text{metaExp}} \rightarrow T_{\text{Exp}}, \\ & \text{followed by } c^*.\end{aligned}$$

The training objective combines classification loss with next-token prediction losses:

$$\begin{aligned}\mathcal{L}_{\text{total}} = & \mathcal{L}_{\text{class}}(c^*, \hat{c}) \\ & + \mathcal{L}_{\text{NTP}}(T_{\text{metaClue}}, T_{\text{Localization}}) \\ & + \mathcal{L}_{\text{NTP}}(T_{\text{metaExp}}, T_{\text{Exp}}),\end{aligned}$$

where $\hat{c}$ is the model's predicted label, $\mathcal{L}_{\text{class}}$ is the classification loss, and $\mathcal{L}_{\text{NTP}}$ denotes the next-token prediction loss for the reasoning components. The hyperparameters λ_1 and λ_2 control the trade-off between classification and reasoning supervision.

We hypothesize that incorporating such structured CoT supervision enables the model to better interpret metaphorical and culturally nuanced content, ultimately improving its performance on complex multimodal reasoning tasks.

F.2 Experimental Setup

Evaluation Metrics: For the multi-class misogyny detection task applied to both our prompt-based and fine-tuned models we report per-class accuracy on each of the four labels (Stereotype, Objectification, Shaming, and Violence) to ensure that performance is tracked individually for all categories. To assess the quality of the generated explanations (i.e., metaphor localization and interpretation), we adopt BERTScore (Zhang et al., 2019) and the LAVE (Mañas et al., 2024) metric.

Implementation Details: We fine-tuned all models using a single NVIDIA A100-SXM4-80GB GPU. Low-Rank Adaptation (LoRA) (Hu et al., 2022) was adopted with configuration parameters: α and r = 64, a dropout rate of 0.05, a learning rate of $2e^{-4}$ and batch size 32. Each model was trained for 4 epochs. We employed LLaMA-Factory (Zheng et al., 2024) for LoRA fine-tuning and VLLM (Kwon et al., 2023) for inference. For reproducibility, we used greedy decoding with

temperature = 0 and no sampling during evaluation. For the other hyperparameters, we followed (Fahim, 2023) configurations.

G Additional Experimental Results

G.1 Fine-Tuning vs. Prompt-Based Performance

Figure 4 illustrates the comparative accuracy (%) of various models across three evaluation paradigms—Zero-Shot, CoT prompting, and LoRA-based Fine-Tuning (LoRA-FT)—on the task of misogyny category classification. In this setup, we treat the prompt-based results (Zero-Shot and CoT) as baselines to assess the impact of task-specific fine-tuning.

Across all model families evaluated—LLaVA, LLaMA, Qwen, Gemma, and PaLI-Gemma—fine-tuned models consistently outperform their prompt-based counterparts. Notably, Qwen 2.5 VL exhibits a substantial improvement, achieving 51.16% accuracy after fine-tuning (LoRA-FT), up from its best prompt-based performance of 39.14% in the Zero-Shot setting. Similarly, Gemma-3, which shows strong performance across all configurations, benefits from fine-tuning, reaching 51.74% accuracy compared to 47.80% (Zero-Shot) and 41.64% (CoT), emerging as the top-performing model overall.

Comparable trends are also observed for LLaVA, LLaMA, and PaLI-Gemma, all of which demonstrate clear performance gains after fine-tuning. While fine-tuning boosts performance over prompt-based approaches, open models still lag behind proprietary closed-source systems in overall accuracy.

G.2 Detailed Analysis of Fine-Tuning Strategies

Limited Gains from Augmented Fine-Tuning: Contrary to expectations, incorporating augmented samples during fine-tuning $LoRA_{aug}$ does not lead to systematic improvements. For example, from the table 4 we can see for LLaMa-3.2, the average performance drops significantly in the $LoRA_{aug}$ setting (25.94%) compared to the $LoRA_{std}$ (45.96%). This pattern is consistent across other models such as Gemma-3 (30.80% vs. 44.69%) and LLaVa-1.5 (25.09% vs. 35.50%), indicating that naïve augmentation strategies may introduce noise or distract the model from core discriminative features.

Effectiveness of Chain-of-Thought Supervised Fine-Tuning: In contrast, figure 4 clearly visualize fine-tuning models using $LoRA_{CoT}$ leads to noticeable improvements over both baseline strategies. For instance, Gemma-3 achieves an average accuracy of 51.74% under $LoRA_{CoT}$—surpassing both its $LoRA_{std}$ (44.69%) and $LoRA_{aug}$ (30.80%) settings. Similar trends are observed with Qwen2.5-VL (51.16% with $LoRA_{CoT}$ vs. 42.59% $LoRA_{std}$), and LLaMa-3.2, which jumps from 45.96% $LoRA_{std}$ to 49.86% $LoRA_{CoT}$. The class-wise breakdown further supports the benefits of CoT supervision. For example, LLaVa-1.5 shows significant gains in the Objectification (52.98%) and Violence (54.26%) categories under $LoRA_{CoT}$, outperforming both $LoRA_{std}$ and $LoRA_{aug}$. These observations underscore the value of reasoning-augmented supervision: by training on rationales, models likely learn to internalize intermediate steps and contextual cues that enhance classification robustness—especially in cases involving subtle or overlapping definitions of misogyny.

While CoT fine-tuning improves overall accuracy and reasoning quality, models still exhibit notable confusion between closely related categories—particularly Shaming, Objectification, and Stereotype. A detailed quantitative analysis of these misclassifications is provided in Appendix G.3.

G.3 Case Studies and Error Analysis

Qualitative Analysis on Metaphor Explanation: Figure 8 shows several examples in which our CoT-fine-tuned models correctly classify the misogyny category while also generating accurate and contextually grounded metaphor explanations. The responses from finetune models such as Gemma-3 and Qwen-2.5 closely align with the gold annotations. We can observe that these models successfully identify problematic content categorized as Stereotype, Objectification, and Violence in memes and images. The models demonstrate sophisticated reasoning by recognizing metaphorical clues (such as the red pill/blue pill analogy), identifying dehumanizing comparisons, and detecting oversimplified gender portrayals. For instance, Gemma-3 correctly interprets the cultural reference to "The Matrix" movie, understanding how the red pill/blue pill dichotomy is being repurposed to create a harmful contrast between different expressions of feminism. Paligemma-2 exhibits spatial reasoning by noting how visual elements (covered vs. revealed

Ground Truth Explanation: The meme uses the Red Pill/Blue Pill metaphor to contrast different expressions of feminism, implying that a Hijabi/Niqabi feminist is less authentic.
Label: Stereotype

Gemma-3: The image uses the metaphor of the red pill and blue pill from the movie "The Matrix," where choosing the red pill leads to harsh reality and the blue pill to blissful ignorance. In this context, the "red pill" is associated with a "feminist" who is visible, while the "blue pill" is linked to a "Hijabi/Niqabi Feminist" who is metaphorically hidden or operating under a "mask." This suggests a negative view of feminists who wear hijab or niqab, implying they are disingenuous or hiding their true beliefs, while those who are more visible are seen favorably. The answer is Stereotype

Paligemma-2: The meme presents two pills, one labeled "Feminist" and the other labeled "Hijabi/Niqabi Feminist". The "Feminist" pill is associated with a hand revealing it publicly, while the "Hijabi/Niqabi Feminist" pill is associated with a hand covering a face, implying that this type of feminism is more private or hidden compared to feminism. The answer Stereotype

Ground Truth Explanation: The image shows a cup designed in the shape of a woman's breast, playing on the sexualized nature of the shape to create a crude objectification of the female body.
Label: Objectification

LLaVA-1.5: The meme holds a double meaning. The image shows a woman holding a coffee mug with exaggerated breasts, while the text makes a sexual innuendo. This reduces the woman to her body parts and sexualizes her by implying something Ironic. The answer is Objectification

Ground Truth Explanation: The meme uses the Joker's dialogue to dismiss the significance of the Shahbag movement and those involved, including women, by comparing their lives to that of a monkey, dehumanizing and belittling their concerns.
Label: Violence

LLaMA-3.2: The text poses a scenario where the Joker claims he will kill 100 people and 1 monkey. The interrogator asks why the monkey. The Joker responds by saying no one cares about 100 Shahbagis, implying that the monkey is a symbol of something insignificant compared to human life.. The answer is violence

Ground Truth Explanation: The top panel depicts a Dog with feminine features labeled "Feminist" and text saying, "If there is income, no need for a husband." The bottom panel shows a cloaked figure with a sword labeled "Muslimah" and text saying, "If there is a husband, no need for income." This sarcastically contrasts two perceived stereotypes: the independent feminist who rejects marriage for financial autonomy, and the traditional Muslim woman who relies on her husband for income, potentially mocking both groups with oversimplified and arguably negative portrayals
Label: Stereotype

Qwen-2.5: The meme sarcastically contrasts the ideas of feminists and Muslims regarding relationships and work. It suggests that feminists want to work but don't want a husband, while Muslim women want a husband but don't want to work. The meme uses these perceptions to mock both groups. The answer is Stereotype.

Figure 8: A qualitative assessment of metaphor explanation generation quality from various finetune models. Given the core instruction to identify the misogyny category with provided options model learns to generate metaphorical explanation before predict the misogyny type. Detailed and accurate parts are emphasized in colored and bold letter.

Ground Truth Explanation: In the meme, foot on neck and touching the sky metaphorically indicate sexual act from different prespective of boy and girl
Label: Shaming

LLaMA-3.2: The meme shows a man irritated by a woman singing along to the radio. The woman is portrayed as annoying and unwanted, like a fungus growing on a foot, with her presence making the man uncomfortable. The answer is Objectification.

LLaVA-1.5: In the meme, cassette used as metaphor for cherishing moment by the couple which indirectly insult women. The answer is Objectification

Ground Truth Explanation: In the meme, tiktok is used metaphorically for which coach should be changed
Label: Shaming

Qwen-2.5: The meme sarcastically implies that the coach's permit is being revoked because he allows the team to use TikTok, suggesting that TikTok usage is detrimental to their performance or-export. The answer is Shaming.

Ground Truth Explanation: The meme mocks women who advocate for sexual freedom in one country but expect conservatism in another, using the slap as a metaphor for silencing perceived hypocrisy
Label: Shaming

Pali-Gemma-2: The meme mocks the idea of playing favorites among friends, creating a sarcastic and relatable scenario. The answer is Shaming

Ground Truth Explanation: The meme uses the wolf to mock feminists and Shahbagis, portraying them as unclean and undesirable, promoting a derogatory and dismissive view of these groups.
Label: Shaming

Gemma-3: The meme uses the image of a bear to suggest that feminists and Shahbagis are so attracted to women that they are a source of constant temptation, implying a negative view of these groups.. The answer is Objectification.

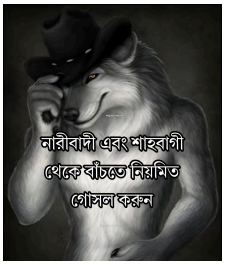


Figure 9: Model Failures in Metaphorical Interpretation of Misogyny detection. Examples illustrating how finetuned models misinterpret metaphors, hallucinate non-existent elements, and generate plausible but incorrect explanations even when occasionally classifying correctly. The red colored is used to highlight the mistakes it made.

hands) reinforce stereotypical implications about different feminist identities. In the case of Objectification, LLaVA-1.5 demonstrates an understanding of how everyday objects can be designed to sexualize female anatomy, recognizing the breast-shaped cup as reducing women to body parts. When analyzing Violence, LLaMA-3.2 correctly identifies the dehumanizing rhetoric that compares human lives (Shahbagis) to animals, understanding how this linguistic device diminishes the value of human life and normalizes violence. Similarly, Qwen-2.5 shows awareness of cultural stereotypes by identifying the oversimplified binary portrayal of feminist and muslim women, recognizing how the meme uses sarcasm to reinforce harmful generalizations about both groups' values regarding work and relationships. Their explanations reveal multi-step analytical processes, first describing the visual and textual elements, then interpreting the implicit messages, and finally connecting these interpretations to the appropriate misogyny categories. This suggests that CoT-based fine-tuning—when grounded in high-quality, richly annotated data—is highly effective in improving both the interpret ability and reasoning ability of vision-language models in socially sensitive tasks such as misogyny detection.

In contrast, misinterpretation of metaphor often results in incorrect classifications. Figure 9 presents several failure cases in which the models demonstrate significant limitations by misclassifying harmful categories and constructing incorrect explanations, even when occasionally arriving at the right label. Key errors include: completely misinterpreting sexual metaphors (such as "foot on neck"), inventing non-existent elements in the images, replacing actual visual components with fabricated ones (e.g., substituting bears for wolves), and creating entirely fictional scenarios unrelated to the actual content. Even when models correctly identify the harmful category, their explanations often fail to capture the intended metaphorical meaning or cultural context, revealing a disconnect between classification accuracy and genuine understanding. These failures highlight critical weaknesses in the models' ability to consistently interpret implicit meanings, recognize culturally-specific references, and reliably distinguish between different categories of harmful content.

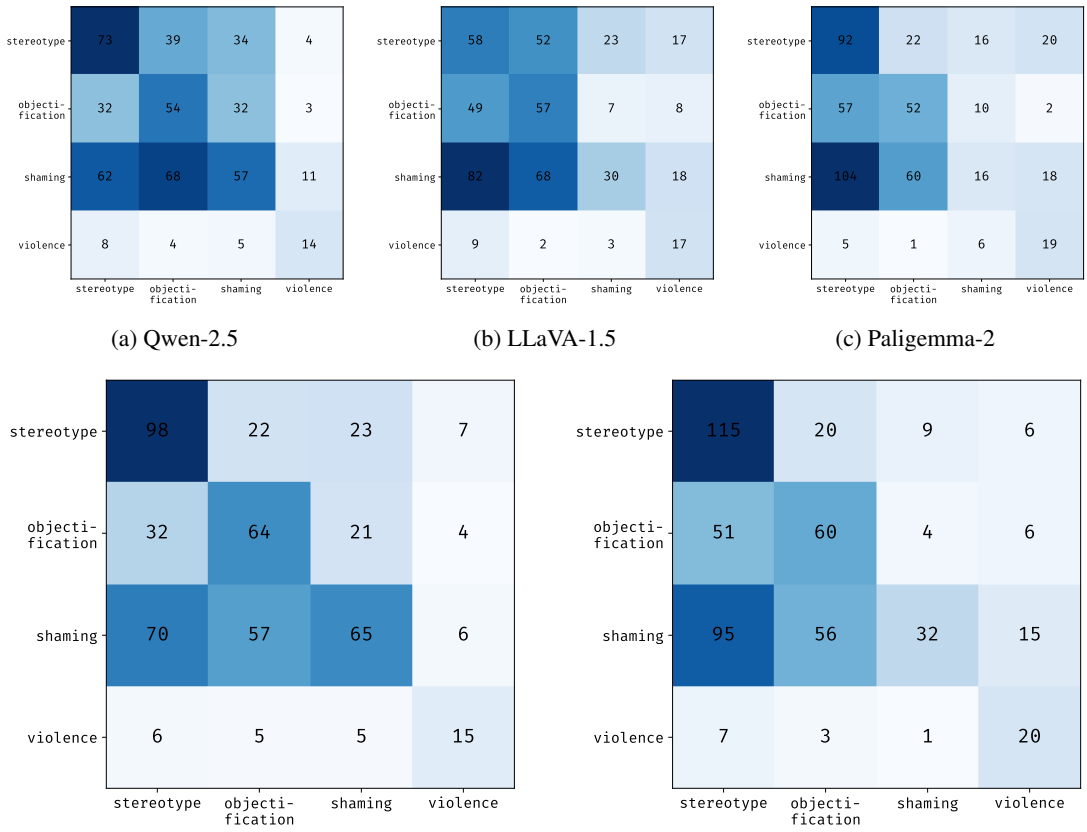
Quantitative Error Analysis: Category Confusion in Misogyny Classification Figure 10 shows that across all models, a consistent pattern emerges where the categories of Shaming and Objectification are frequently confused with each other and with Stereotype. This classification confusion reveals significant challenges in distinguishing between different forms of misogynistic content. For example in the Gemma-3 model, only 16.16% of Shaming instances are correctly classified 95 out of 198 Shaming instances are misclassified as Stereotype 56 out of 198 Shaming instances are misclassified. As Objectification for the Objectification class, while 49.59% of instances are correctly identified 51 out of 121 Objectification instances are wrongly labeled as Stereotype. This is the consistent pattern we can see from all the models. The pattern clearly indicates that the Shaming class is consistently misclassified, especially as Stereotype and Objectification, suggesting that models struggle to distinguish between Shaming and other misogyny categories. Hence, this remains an open question: what underlying conceptual overlaps exist between these misogyny categories, and how might we develop more nuanced taxonomies and training approaches to help models better differentiate among them?

G.4 Performance of VLMs on Humor Type

Table 7 shows the performance of various VLMs in detecting five types of humor: Ironic, Mockery, Satirical, Sarcastic, and Other across three prompting methods: Zero Shot, CoT, and LoRA (CoT) Fine-Tuning. Closed Source VLMs like Gemini2.0 Flash and GPT-4o-mini generally perform better than open-source models across humor types. Gemini2.0 Flash consistently excels, with an average score of 47.15 in Zero Shot and 50.78 in CoT. GPT-4o-mini performs slightly worse, especially under CoT (avg: 42.70). Among open-source models, Gemma-3-12B leads with an average of 47.34 in Zero Shot, outperforming others like Qwen2.5-VL 7B (avg: 45.45) and Llama-3.2V 11B (avg: 29.01).

Performance by humor type shows that Gemini2.0 Flash excels at detecting Ironic, Mockery, and Sarcastic humor, while Gemma-3-12B is strong in Mockery and Satirical humor. Qwen2.5-VL 7B stands out in the Other humor category with a notably high score of 70.11 in Zero Shot. On the other hand, LLaVA-1.5 7B and LLaVA-NeXT 7B consistently score low across all humor types, especially in categories like Satirical and Sarcastic humor.

When comparing the prompting techniques, Zero Shot results show Gemini2.0 Flash and Gemma-3-12B as top performers, while LLaVA



(a) Qwen-2.5 (b) LLaVA-1.5 (c) Paligemma-2

(d) LLaMA-3.2 (e) Gemma-3

Figure 10: Confusion matrices for all finetuned models showing category-wise classification performance on the misogyny detection task. Rows represent the true labels and columns the predicted labels. Frequent confusion is observed between shaming and objectification.

models perform poorly. With Chain of Thought (CoT) prompting, Gemini2.0 Flash shows improvement, while GPT-4o-mini experiences a slight decline. In LoRA Fine-Tuning, Gemma-3-12B and Qwen2.5-VL 7B see a performance boost, particularly in Mockery and Sarcastic humor, while LLaVA-1.5 7B shows minimal improvement.

Overall, Gemini2.0 Flash remains a standout model, especially in Ironic and Sarcastic humor detection, while LLaVA models show consistent weaknesses.

G.5 Misogyny Intensity and Metaphor Localization Results

Table 8 presents the performance of different VLMs in detecting Misogyny Intensity (high, moderate, low) and Metaphor Localization (text, image, both) across three prompting techniques: Zero Shot, Chain of Thought (CoT), and LoRA (CoT) Fine-Tuning.

In Zero Shot Prompting, Gemini2.0 Flash emerges as a strong performer in both Misogyny Intensity (avg: 49.41) and Metaphor Localization (avg: 50.07), consistently outperforming GPT-4o-mini (avg: 47.82 and 48.08, respectively). Among open-source VLMs, Gemma-3-12B demonstrates solid performance with an average of 33.78 for Misogyny Intensity and 37.57 for Metaphor Localization. Other models like Llama-3.2V 11B and LLaVA models score significantly lower, especially in Metaphor Localization where they lag behind, with LLaVA-NeXT 7B scoring the lowest in both categories (avg: 17.19 for Misogyny Intensity and 20.04 for Metaphor Localization).

Under Chain of Thought (CoT) prompting, Gemini2.0 Flash again leads with an average of 53.66 for Misogyny Intensity and 54.11 for Metaphor Localization, showing a noticeable improvement over its Zero Shot performance. GPT-4o-mini maintains consistent results with a slight increase in Misogyny Intensity (avg: 47.40) and a small decline in Metaphor Localization (avg: 48.06). Among open-source models, Gemma-3-12B continues to show strength in Metaphor Localization (avg: 40.46) but

Models	Humor Types					
	Ironic	Mockery	Satirical	Sarcastic	Other	Avg
Zero Shot Prompt						
Closed Source VLMs						
Gemini2.0 Flash	44.07	51.37	51.72	51.76	36.84	47.15
GPT-4o-mini	51.61	50.25	50.00	46.56	30.00	45.68
Open Source VLMs						
Llama-3.2V 11B	22.35	30.12	32.68	29.47	30.83	29.01
Gemma-3-12B	45.27	52.13	50.94	45.76	42.58	47.34
Qwen2.5-VL 7B	50.62	40.18	25.37	40.95	70.11	45.45
Phi-3.5	38.03	35.64	18.92	34.87	30.09	31.51
Phi-4	36.41	34.08	15.73	33.21	30.64	30.01
LLaVA-1.5 7B	35.78	32.29	17.85	28.46	12.57	25.40
LLaVA-NeXT 7B	28.12	27.64	15.31	27.88	25.39	24.87
Chain of Thought (CoT) Prompt						
Gemini2.0 Flash	49.15	51.63	55.17	57.89	36.84	50.78
GPT-4o-mini	47.37	51.91	35.71	45.18	33.33	42.70
Open Source VLMs						
Llama-3.2V 11B	16.67	23.94	27.27	23.19	25.00	23.22
Gemma-3-12B	36.67	46.46	42.86	33.33	30.00	37.66
Qwen2.5-VL 7B	40.00	32.08	17.65	33.64	62.50	37.57
Phi-4	26.67	28.67	8.33	27.92	25.00	23.52
LLaVA-1.5 7B	26.19	26.53	9.52	20.28	6.67	17.84
LLaVA-NeXT 7B	21.43	20.77	10.34	20.69	20.00	18.85
LoRA (CoT) Fine-Tuning						
Llama-3.2V 11B	35.29	41.67	43.10	39.72	38.24	39.01
Gemma-3-12B	48.78	53.33	51.28	47.50	44.44	49.07
Qwen2.5-VL 7B	47.94	44.23	37.50	42.86	68.18	48.76
LLaVA-1.5 7B	41.18	38.10	29.17	33.33	20.00	32.75
Paligemma-2-10B	45.01	49.12	42.34	43.71	38.18	43.07

Table 7: Performance of different VLMs in detecting various ***Humor Types*** under Zero Shot, Chain of Thought (CoT) prompting, and LoRA fine-tuning.

lags behind in Misogyny Intensity (avg: 39.60). Qwen2.5-VL 7B also performs reasonably well in Misogyny Intensity (avg: 34.09) and Metaphor Localization (avg: 30.91).

In LoRA Fine-Tuning, Gemma-3-12B continues to outperform most other models, achieving an average of 45.24 for Misogyny Intensity and 49.36 for Metaphor Localization, indicating that fine-tuning improves its performance. Similarly, Qwen2.5-VL 7B shows notable improvements (avg: 40.37 and 38.83, respectively). Llama-3.2V 11B, while showing a slight improvement in Misogyny Intensity (avg: 28.53), does not perform as well in Metaphor Localization (avg: 26.91). LLaVA-1.5 7B and Paligemma-2-10B also show improvements, with Paligemma-2-10B being the top performer among fine-tuned models (avg: 46.28 for Misogyny Intensity and 48.34 for Metaphor Localization).

Overall, Gemini2.0 Flash consistently excels across all categories and prompting methods, while Gemma-3-12B shows the most improvement under fine-tuning, especially in Metaphor Localization. The LLaVA models struggle, particularly in Metaphor Localization, highlighting a clear performance gap between closed-source and open-source models.

G.6 Meme Template based Result Analysis

Table 9 compares the performance of different VLMs on meme template classification, assessing results across Zero Shot, Chain of Thought (CoT), and LoRA (CoT) Fine-Tuning. The models are evaluated on various meme templates such as Non-Templated, Troll Face, Wojak, Rage Comic, Soyjak, Doge, Other, and their overall average.

Under Zero Shot Prompting, Gemini2.0 Flash shows strong results, particularly with Doge (96.02) and Troll Face (63.64), but its average score

Models	Misogyny Intensity				Metaphor Localization			
	High	Moderate	Low	Avg	Text	Image	Both	Avg
Zero Shot Prompt								
Closed Source VLMs								
Gemini2.0 Flash	53.93	45.33	49.58	49.41	52.05	55.91	43.24	50.07
GPT-4o-mini	53.33	45.31	43.83	47.82	47.70	50.00	47.80	48.08
Open Source VLMs								
Llama-3.2V 11B	22.35	20.12	13.68	18.05	21.47	20.83	15.24	19.34
Gemma-3-12B	36.27	34.13	30.94	33.78	31.76	44.58	36.38	37.57
Qwen2.5-VL 7B	27.62	30.18	34.37	30.39	34.95	32.11	23.41	30.82
Phi-3.5	18.03	23.64	30.92	24.53	25.87	30.09	21.04	25.75
Phi-4	17.41	22.08	28.73	22.07	23.21	28.64	18.61	24.92
LLaVA-1.5 7B	19.78	18.29	22.85	20.64	18.46	22.17	14.57	18.73
LLaVA-NeXT 7B	17.12	17.64	19.31	18.02	15.88	18.39	17.09	17.19
Chain of Thought (CoT) Prompt								
Closed Source VLMs								
Gemini2.0 Flash	54.97	48.34	56.67	53.66	52.27	55.91	53.02	54.11
GPT-4o-mini	49.17	44.97	47.06	47.40	46.15	56.32	43.26	48.06
Open Source VLMs								
Llama-3.2V 11B	27.63	23.08	15.56	22.42	27.71	22.50	16.00	20.73
Gemma-3-12B	40.45	42.86	34.48	39.60	34.82	50.00	41.56	40.46
Qwen2.5-VL 7B	30.19	33.33	38.75	34.09	37.12	36.54	26.97	30.91
Phi-4	20.83	27.64	35.71	28.06	28.42	34.62	19.01	24.92
LLaVA-1.5 7B	21.52	20.83	25.56	22.64	22.99	27.16	17.70	22.57
LLaVA-NeXT 7B	19.70	20.27	20.69	20.22	18.92	23.08	20.13	20.04
LoRA (CoT) Fine-Tuning								
Llama-3.2V 11B	32.81	29.04	22.73	28.53	31.13	29.18	21.45	26.91
Gemma-3-12B	46.93	48.21	40.57	45.24	42.59	56.25	48.92	49.36
Qwen2.5-VL 7B	36.84	39.37	43.89	40.37	42.16	41.24	33.09	38.83
LLaVA-1.5 7B	27.78	26.12	30.77	28.56	29.67	33.45	23.19	28.31
Paligemma-2-10B	45.51	46.73	47.61	46.28	45.92	52.42	45.67	48.34

Table 8: Performance of various VLMs across ***Misogyny Intensity and Metaphor Localization Categories*** under Zero Shot and Chain of Thought prompting, and LoRA fine-tuning.

is 44.69, which lags behind GPT-4o-mini (avg: 50.23). The open-source VLMs show relatively lower performance across the board. For instance, Gemma-3-12B achieves the highest average among the open-source models at 41.96, with reasonable performance in Troll Face (39.39) and Other (40.10). Llama-3.2V 11B performs slightly better than others with an average of 38.11, with Doge (94.23) being its standout.

In Chain of Thought (CoT) prompting, Gemini2.0 Flash again leads with an average of 47.51, with notable improvement in Non-Templated (54.92) and Doge (95.00). GPT-4o-mini also shows a solid performance with an average of 45.87, especially excelling in Wojak (50.00) and Doge (95.05). Among open-source models, Llama-3.2V 11B performs lower than the closed-source models, with an average of 30.48. Gemma-3-12B and Qwen2.5-VL 7B achieve comparable averages of 36.98 and 34.79, respectively, with Gemma showing decent performance in Soyjak (10.22).

Under LoRA (CoT) Fine-Tuning, the models show mixed results. Llama-3.2V 11B reaches an average of 34.83, with strong performance in Doge (96.34). Gemma-3-12B maintains consistent performance with an average of 36.88, especially in Troll Face (37.45) and Other (44.36). Qwen2.5-VL 7B shows a slightly lower average of 34.83 but does well in Doge (82.47) and Other (42.71). LLaVA-1.5 7B, while performing decently in Doge (97.12), achieves a lower average of 33.57, with notable underperformance in Non-Templated (22.28). Paligemma-2-10B, a strong performer across categories, achieves an average of 38.53, excelling in Doge (89.88).

Gemini2.0 Flash consistently outperforms all

Models	Meme Template							
	Non-Templated	Troll Face	Wojak	Rage Comic	Soyjak	Doge	Other	Avg
Zero Shot Prompt								
Closed Source VLMs								
Gemini2.0 Flash	50.13	63.64	33.33	10.35	33.33	96.02	48.65	44.69
GPT-4o-mini	48.57	50.00	40.00	33.33	75.00	95.00	41.03	50.23
Open Source VLMs								
Llama-3.2V 11B	30.45	41.67	18.20	15.45	38.89	94.23	22.76	38.11
Gemma-3-12B	42.17	39.39	31.11	30.55	15.56	84.90	40.10	41.96
Qwen2.5-VL 7B	36.88	36.36	20.00	27.78	19.75	82.12	38.65	36.88
Phi-3.5	32.12	34.00	23.08	20.00	22.22	84.78	35.33	31.54
Phi-4	28.55	38.89	25.64	18.12	21.67	86.42	37.50	33.71
LLaVA-1.5 7B	25.75	31.11	35.00	17.35	27.78	95.88	31.02	34.38
LLaVA-NeXT 7B	23.60	20.00	37.04	14.78	30.00	89.33	28.89	31.04
Chain of Thought (CoT) Prompt								
Closed Source VLMs								
Gemini2.0 Flash	54.92	45.45	22.22	33.33	33.33	95.0	50.00	47.51
GPT-4o-mini	47.59	45.00	50.00	30.35	25.00	95.05	47.22	45.87
Open Source VLMs								
Llama-3.2V 11B	23.36	33.33	10.22	12.33	50.00	95.67	13.64	30.48
Gemma-3-12B	40.11	36.36	28.57	33.33	10.22	80.23	43.48	36.98
Qwen2.5-VL 7B	34.06	33.33	12.20	33.33	15.30	80.85	41.67	34.79
Phi-4	24.92	42.11	28.57	10.25	15.71	81.86	40.62	34.68
LLaVA-1.5 7B	21.15	26.67	40.00	11.20	25.00	96.20	28.57	32.87
LLaVA-NeXT 7B	19.79	10.00	44.44	11.36	33.33	87.50	25.00	30.53
LoRA (CoT) Fine-Tuning								
Llama-3.2V 11B	24.90	34.66	10.53	12.95	51.50	96.34	14.05	34.83
Gemma-3-12B	41.31	37.45	29.43	34.33	10.53	81.84	44.36	36.88
Qwen2.5-VL 7B	35.75	34.33	12.67	34.33	15.80	82.47	42.71	34.83
LLaVA-1.5 7B	22.28	27.47	41.20	11.64	25.75	97.12	29.42	33.57
Paligemma-2-10B	40.12	38.51	30.13	30.10	28.90	89.88	42.20	38.53

Table 9: Performance of various VLMs on meme templates under Zero Shot, Chain of Thought prompting, and LoRA fine-tuning.

models in Zero Shot and Chain of Thought prompting, with GPT-4o-mini closely following. Llama-3.2V 11B and Gemma-3-12B show solid performances in various meme templates, but open-source VLMs generally lag behind the closed-source models in terms of average scores. Fine-tuning with LoRA improves the performance of most models, with Llama-3.2V 11B and Paligemma-2-10B being the most notable models in the fine-tuned setting.

H Used Prompts in the Paper

Prompt Used for LAVE Evaluation

Prompt for LAVE evaluation

You are an expert cultural anthropologist tasked with evaluating the correctness of candidate answers for cultural visual question-answering.
Given an image as context, a question, and reference answer by an expert, and a candidate answer by a model, please rate the candidate answer's correctness.
Use a scale of 0-1, where 0 indicates an incorrect, irrelevant, or imprecise answer, and 1 indicates a correct, precise answer according to the reference.
You have to provide the rationale for your rating and then provide a rating in a specific '*rating:* X' format, where X is either 0 or 1.

ZeroShot Prompt.

ZeroShot Prompt for Misogyny Detection

You are an expert at detecting misogynistic content in Bangla memes (containing both image and text, including text written in Latin script/English letters). Your task is to classify memes into one of four categories of misogyny: Stereotype, Objectification, Shaming, or Violence, and provide a very brief explanation.

Chain-of-Thought Prompt.

CoT Prompt for Misogyny Detection

You are an expert in analyzing Bangla memes for misogynistic content. Think through each step carefully before answering.
Step 1: Briefly describe what you see in the image (scene, characters, actions).
Step 2: Transcribe and interpret the text, including any Bangla-English code-mixing or slang.
Step 3: Note any cultural references that influence the meme's meaning.
Step 4: Identify the humor type:
– Mockery (ridicule),
– Sarcasm (opposite of what is meant),
– Irony (contrast between expectation and reality),
– Satire (criticism via humor),
– Other (specify).
Step 5: Check for metaphor—if present, explain where (image/text) and its meaning.
Step 6: Select the most relevant misogyny category:
– Stereotype: Content that attributes generalized characteristics to women, reinforcing traditional gender roles.
– Objectification: Content that treats women as objects devoid of agency, focusing solely on their physical appearance.
– Shaming: Content that criticizes, mocks, or ridicules women for their appearance, behaviors, choices, sexuality, or for not conforming to gender norms.
– Violence: Content that promotes, normalizes, or makes light of physical, sexual, or emotional violence against women.
Step 7: Write a 1–2 sentence explanation of the misogynistic content.
Step 8: Present the final answer in this format:
CLASSIFICATION: [Category]
MEME EXPLANATION: [Explanation from Step 7]"""