On VLMs for Diverse Tasks in Multimodal Meme Classification

Anonymous ACL submission

Abstract

The use of multimodal memes to spread hatred, propaganda, violence, in social and digital media can be tackled by content moderation by AI-based meme analysis. In this paper, we present a comprehensive and systematic analysis of vision-language models (VLMs) for disparate meme classification, and introduced a novel approach **Co**mbining **V**LM **Ex**planation to **Fi**ne-tune **L**LMs (CoVExFiL). In the proposed CoVExFiL, we generated a VLM-based understanding of meme images and fine-tuned the LLMs on textual understanding of the embedded meme text. Our contributions are threefold: (1) Benchmarking VLMs with diverse prompting strategies for these sub-tasks; (2) Evaluating LoRA fine-tuning across all VLM components to assess performance gains; and (3) Proposing a novel approach where detailed meme interpretations generated by VLMs are used to train smaller language models (LLMs), significantly improving classification. After extensive experimentation, we observe that CoVExFiL improved the baseline performance by 8.34%, 3.52% and 26.24% for sarcasm, offensive and sentiment classification, respectively. These findings shed light on the capabilities and shortcomings of VLMs, while also establishing CoVExFiL as a promising strategy for advancing meme understanding.

1 Introduction

Multi-modal memes have gained popularity (Petrova, 2021) due to their eloquent and powerful way to convey complex, subtle messages (Das, 2023). The widespread use of memes propagating hatred (Gelber and McNamara, 2016), abuse, misogyny, and propaganda poses a serious concern (Bhattacharya, 2019) that can be resolved by AI-based meme understanding (Hee et al., 2024). AI-based meme understanding can facilitate social media moderation, digital content filtering, hate speech mitigation, and even early detection of online radicalization. In an attempt to resolve this issue, the AI community has published several useful datasets and model frameworks, as detailed in Appendix A.

These studies explored the capabilities of multimodal deep learning models, Large Language Models (LLMs) and Vision-Language Models (VLMs) (Afridi et al., 2021). While traditional deep learning models show limited generalization across different meme types (Shah et al., 2024), the LLMs excel at text processing but are incapable of processing visual information (Cai et al., 2025). On the contrary, the VLMs demonstrate strong multimodal reasoning capabilities. However, the performance of these VLMs varies significantly based on contextual understanding (Xing et al., 2024). The existing approaches achieve subpar efficacy, probably because they cannot comprehend the complementary and indirect linkage between text and image(Zhong and Baghel, 2024).

Although VLMs are capable of rich, context-aware multi-modal reasoning, they require substantial computing resources for fine-tuning (Zhang et al., 2024; Li et al., 2025). Additionally, most VLMs are trained on general image-text datasets without meme-specific content, which constrains their ability to develop specialized understanding of this domain. Besides, accurate meme understanding necessitates relating the information to the contextual understanding and the cultural backdrop (Yus, 2019). Given these challenges, achieving accurate meme classification poses considerable difficulties for existing AI architectures (Jha et al., 2024). In contrast, lightweight LLMs such as BERT show strong performance on text-based classification tasks, including sentiment analysis, emotion detection, and offensive language identification (Devlin et al., 2019; Sosea and Caragea, 2025; Barbieri et al., 2020). The main limitation lies in the trade-off between capability and efficiency: VLMs can understand multi-modal content well (Caffagni et al., 2024) but are resource-

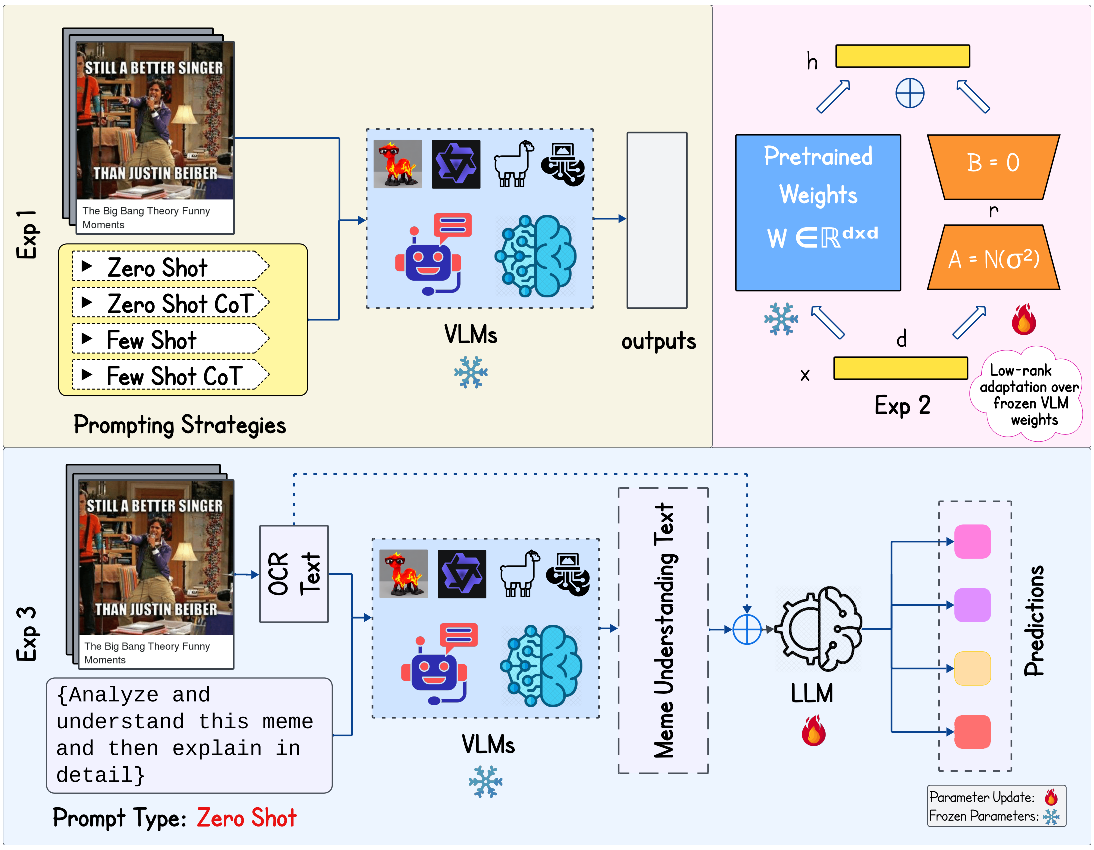


Figure 1: Schematic diagram of the three strategies mentioned in the experiment section.

intensive to train, while lightweight LLMs are efficient and show effectiveness for text-based classification tasks (Chang et al., 2024) but cannot process images. This contrast opens up an opportunity to combine their strengths for more effective meme classification by leveraging their complementary capabilities. By combining high-level reasoning from VLMs with LLMs through textual representations, we can achieve three key benefits. First, we can leverage the multi-modal understanding capabilities of the VLMs, leading to a significant reduction of computational overhead for deployment compared to fine-tuning VLMs directly. Second, we can provide rich training data in the form of VLM-generated textual understandings that can be fed directly to LLMs. Third, we can efficiently fine-tune LLMs to classify the memes for downstream tasks, including sentiment, humor, sarcasm, offense, motivation, and misogyny detection.

We utilized widely used open-sourced VLMs in different settings as illustrated in Fig. 1 to address the following research questions-

- **R1:** Are the pre-trained VLMs with diverse types of prompting powerful enough to understand memes?

- **R2:** Does fine-tuning VLMs using adapters improve the performance?

- **R3:** Can we combine pre-trained VLMs contextual understanding to train the LLMs for better classification?

The important contributions of the work are:

- **Benchmarking VLM Meme Classification:** *We benchmark the performance of various VLMs using diverse prompting strategies such as Zero Shot (ZS), Zero Shot Chain-of-Thought (ZSC), Few Shot (FS), Few Shot Chain-of-Thought (FSC).*

- **Evaluating LoRA Fine-Tuning:** *We systematically explore the parameter-efficient LoRA fine-tuning approach on diverse meme understanding tasks.*

- **Novel Meme Understanding Approach:** *We introduce an innovative approach called Combining VLM Explanation to Fine-tune LLMs (CoVExFiL). In this approach, we use VLM-generated meme interpretations from multiple* *prompting techniques to train LLMs. The proposed approach with a three-step CoT prompting produces the best results.*

Some of the key findings from this paper are as follows-

- Fine-tuning VLMs using the LoRA adapter was comparatively less effective for our task, likely due to its constraint of updating only a small set of additional adapter parameters rather than the full model weights. On the other hand, the Qwen model surpassed the baseline performance for sentiment and sarcasm classification under Few Shot prompting by 16.71%, and 0.39%, respectively.

- The proposed CoVExFiL strategy yields noticeable performance gains in tasks such as sarcasm and offensiveness detection, with a significant 26.10% improvement in SN, outperforming SOTA. This performance gain can be attributed to its strong meme understanding, and error analysis further confirms the effectiveness of training on generated explanations.

2 Datasets and Experimental Setup

We used two popular datasets in our experiments: **Memotion**(Sharma et al., 2020) and **MAMI** (Multimedia Automatic Misogyny Identification)(Fersini et al., 2022). The Memotion dataset categorizes memes into five emotion-related tasks—humor (HM), sarcasm (SR), offensiveness (OF), sentiment (SN), and motivation (MV). Among these, humor, sarcasm, offensiveness, and sentiment are multiclass classification tasks, while motivation is a binary classification task. These tasks span a broad spectrum of meme classification challenges, and to the best of our knowledge, this is the first work that explores all of them collectively in a unified framework. On the other hand, the MAMI dataset consists of two sequential tasks: first, detecting whether a meme is misogynistic (MG), which is formulated as a binary classification task; and second, if misogyny is detected, identifying the specific type(s) of misogyny (MGT) through a multilabel classification task. We have detailed the statistics of the datasets in Table 1. We employ LLAVA, Qwen, LLaMA, and InstructBLIP open-source vision-language models (VLMs) in our study, as detailed in Table 2.

Dataset	Task	Labels	# Samples	Length
Memotion	Sentiment (SN)	Positive	4,160	13.27
		Neutral	2,201	12.85
		Negative	631	13.57
	Humor (HM)	Not Funny	1,651	13.71
		Funny	2,452	12.69
		Very Funny	2,238	13.02
		Hilarious	651	13.09
	Sarcasm (SR)	Not Sarcastic	1,544	13.19
		General	3,507	13.02
		Twisted Meaning	1,532	13.08
		Very Twisted	394	13.10
	Offensive (OF)	Not Offensive	2,713	13.31
		Slight	2,592	13.00
		Very Offensive	1,466	12.74
		Hateful Offensive	221	13.21
	Motivation (MV)	Not Motivational	4,525	13.01
		Motivational	2,467	13.19
MAMI	Misogyny (MG)	Misogynous	5,500	16.21
		Non-Misogynous	5,500	19.90
	Misogyny type (MGT)	Shaming	1,420	18.54
		Stereotype	3,160	18.34
		Objectification	2,550	18.93
		Violence	1,106	18.26

Table 1: Dataset statistics. Here, 'Length' denotes the average caption length in words.

VLM	Size	Version
LLAVA-1.6 (LV)	7B	`llava-hf/llava-v1.6-mistral-7b-hf`
Qwen2-VL (QW)	7B	`Qwen/Qwen2-VL-7B-Instruct`
LLaMA-3.2-Vision (LM)	11B	`meta-llama/Llama-3.2-11B-Vision-Instruct`
InstructBLIP (IB)	7B	`Salesforce/instructblip-vicuna-7b`

Table 2: Selected VLMs and their corresponding versions used in our experiments.

3 Experiments

In this section, we report the details of three types of experiments, each guided by a distinct research philosophy. While the first experiment explores the effectiveness of different prompting strategies, the second experiment finetunes the LoRA adapter associated with the VLMs reducing computational cost. Our main contribution, *CoVExFiL*, lies in the third experiment, which introduces a novel two-step approach for enhanced meme understanding. Details of each experiment are provided in the subsections.

3.1 Experiment-1 : Prompting VLMs using various methods

One effective way to guide model outputs in low-resource and few-shot settings is through prompting, which does not require altering model parameters (Alayrac et al., 2022; Zhou et al., 2024). In this experiment, we evaluated various prompting strategies such as Zero Shot (*ZS*), Zero Shot Chain-of-Thought (*ZSC*), Few Shot (*FS*), and Few Shot Chain-of-Thought (*FSC*) to classify memes. While in the ZS setting we prompted VLMs to classify the memes, in ZSC (Wei et al., 2022; Xu et al., 2024a) we additionally asked the VLMs to provide detailed step-by-step reasoning. In the FS version, we followed similar steps; however, in addition, we included some example input for context. The FSC setting combined the FS approach with CoT-based reasoning. The specific prompts used in each strategy, along with their corresponding experimental settings, are illustrated in Appendix B.3. Figure 5 presents detailed query formats and prompt structures for all four strategies, making it easier to understand the differences and design choices involved. Also, to ensure reproducibility and minimize randomness in responses, we maintained a low temperature setting (0.1) for the VLM throughout all experiments.

3.2 Experiment-2: Fine-tuning VLMs using LoRA adapter

In this experiment, we explored a substitute option of employing LoRA (Low-Rank Adaptation) (Hu et al., 2022), a parameter-efficient approach for fine-tuning VLMs. Instead of updating all model parameters, LoRA trains lightweight adapter layers, enabling VLMs to efficiently adapt to the meme classification task. Since the performance of the pre-trained VLMs depends on the training data, they perform well only if the model is trained on similar data. For LLAVA, Qwen, and LLaMA, we fine-tuned LoRA adapters in all vision and language layers, including the attention and MLP modules, ensuring comprehensive adaptation for both image encoding and text generation. In contrast, for InstructBLIP, we fine-tuned only the *q_proj* and *v_proj* layers, focusing on the core multimodal fusion mechanism to enhance cross-modal alignment with minimal resource usage. For LoRA-based fine-tuning, we used the following hyperparameters a rank of 16 and set the LoRA scaling factor α to 32. The fine-tuning was performed for 2 epochs using the `adamw_8bit` optimizer with a learning rate of 2×10^{-4}.

3.3 Experiment-3: CoVExFil

In our earlier experiments on meme classification, we explored a variety of prompting strategies and used LoRA-based fine-tuning as a computationally efficient alternative to full-scale VLM fine-tuning. Moving a step beyond these methods, i) first, we focused on leveraging the interpretability of VLMs to generate richer textual understandings of memes on the various prompts (Figure 5). Rather than relying solely on final predictions, we aimed to extract these intermediate understandings, and subsequently, (ii) we used them to guide lightweight LLMs for downstream classification. For this purpose, we propose a novel, two-step, interpretable approach, termed CoVExFiL (Combining VLM Explanation to Fine-tune LLMs), where we prompt the LMs to generate textual understanding of the meme image, which is used to fine-tune the LLMs. The key motivation for this approach is to harness the extensive world knowledge that VLMs acquire from their pre-training on massive, diverse datasets, making them uniquely suited to address the inherent diversity and evolving nature of memes. This contrasts with standard fine-tuning on a limited dataset, which often fails to generalize effectively to novel memes (Nguyen and Ng, 2024a). Instead of undertaking the computationally expensive task of fine-tuning the VLM itself, we leverage its capabilities to generate rich data by various prompting, which in turn can efficiently fine-tune much smaller LLMs. We considered three widely popular and powerful pre-trained LLMs: BERT (Devlin et al., 2019), RoBERTa (Liu et al., 2019), and XLNet (Yang et al., 2019)(as detailed in Appendix B.2.1). This experiment aims to investigate whether VLM-generated meme understandings can help and enhance downstream classification capabilities of LLMs. To the best of our knowledge, we are the first to explore this approach across a diverse range of tasks in meme understanding.

MDL	PM	Memotion					MAMI		Avg.
		HM	SR	OF	SN	MV	MG	MGT	
LV	ZS	27.43	13.82	11.57	24.73	25.67	57.31	14.65	25.03
	ZSC	29.83	12.76	24.55	32.35	47.33	68.38	31.06	35.18
	FS	25.94	**32.98**	12.72	16.57	51.42	75.46	34.93	35.72
	FSC	28.30	18.92	23.44	27.91	55.96	**77.65**	28.77	37.28
QW	ZS	25.97	14.82	19.89	35.49	51.58	49.12	13.24	30.02
	ZSC	27.29	12.79	24.88	36.09	**57.62**	71.77	25.55	36.57
	FS	27.57	10.23	**29.41**	**44.70**	53.06	70.87	25.74	37.37
	FSC	**31.34**	22.18	27.72	44.19	52.69	73.31	**38.92**	41.48
LM	ZS	24.76	18.23	20.79	20.32	19.34	51.65	12.37	23.92
	ZSC	29.91	17.85	22.68	41.48	49.80	60.76	27.79	35.75
IB	ZS	22.85	19.47	13.69	11.25	18.12	45.47	11.18	20.29
	ZSC	23.85	28.64	22.69	20.25	45.12	49.47	20.18	30.03
Avg.		27.09	18.56	21.17	29.61	43.98	62.60	23.70	
Std.		±2.56	±6.72	±5.78	±11.17	±14.33	±11.64	±9.29	
SOTA △		49.09	32.85	34.38	38.30	59.28	87.40	73.14	

Table 3: Performance of VLMs via prompting approaches. The highest F1 score for each task is in **blue** (column-wise).

MDL	Memotion					MAMI		Avg.
	HM	SR	OF	SN	MV	MG	MGT	
LV	30.21	17.58	11.57	29.83	49.16	**64.16**	24.53	32.43
QW	**36.35**	20.97	**29.89**	30.34	**51.58**	61.33	**27.83**	36.89
LM	33.95	**23.17**	26.34	**31.72**	49.33	58.98	22.13	35.08
IB	23.03	15.98	18.98	28.61	35.19	47.33	19.83	26.99
Avg.	30.89	19.43	21.70	30.13	46.32	57.95	23.58	
Std.	±4.86	±2.84	±7.52	±1.28	±6.66	±6.50	±3.37	
SOTA △	49.09	32.85	34.38	38.30	59.28	87.40	73.14	

Table 4: LoRA Adapter Fine-Tuning (Experiment 2). The highest F1 score for each task is shown in **blue** (column-wise).

4 Results and Discussion

We quantify the performance using **average weighted F1 score** to counter the class imbalance issue. We compare results task-wise and model-wise, highlight key trends, and show how each strategy enhances meme classification performance.

4.1 Observations for Experiment 1

Table 3 demonstrates that FSC prompting improves performance in all tasks, with QW achieving the highest AWF1 **41.48**. This result highlights the importance of structured reasoning and its effectiveness in enhancing multimodal understanding. Task-wise, misognyny (MG) shows the best performance (**77.65**), likely due to the presence of explicit cues. In contrast, tasks like sarcasm (SR) and offensiveness (OF) resulted in lower average weighted F1 score (**18.56–29.61**) on average, reflecting challenges in contextual and complexity. Model-wise, QW and LV respond best to prompting strategies, with LV attaining **49%** gain from ZS to FSC. In contrast, IB and LM perform poorly under ZS (**20.29**, **23.92**), indicating limited reasoning without guidance. Although prompting lags behind SOTA, three-step CoT shows promise. Notably, in tasks such as misogyny (MG) and motivational (MV), the performance gains remain modest.

4.2 Observations for Experiment 2

Table 4 reporting the efficacy after using LoRA fine-tuning affirms that this method does not lead to noticeable improvement. Tasks such as humor (HM), sarcasm (SR), offensiveness (OF), sentiment (SN), motivation (MV), misogyny (MG), and misogyny type (MGT) report relative improvement of 14.01%, 4.69, 2.50, 1.76, 5.32, -7.43, -0.51, respectively. Motivational (MV) and misogyny (MG) detection tasks attained relatively higher scores, while sarcasm (SR) and offensiveness (OF) remained challenging. Compared to Exp 1, LoRA underperforms in terms of best model performance (e.g., Qwen-FSC: **41.48** in Exp 1 vs. Qwen-LoRA: **36.89**). While LoRA fine-tuning slightly improves task-wise averages, prompting-based methods, especially CoT, consistently obtain better model performance. Overall, Qwen-FSC achieves significantly better results with prompting (e.g., 41.48 vs. 36.89), showing that prompting strategies handle complex meme understanding more effectively than LoRA.

4.3 Observations for Experiment 3

As per the results shown in Table 5, it significantly outperformed prior experiments. CoVExFiL achieved relative gains of 23.10% (HM), 7.92% (SR), 21.01% (OF), 8.07% (SN), 1.86% (MT), and 2.32% (MG) over the best prompting scores (Exp 1). Notably, CoVExFiL reports a **26.14%** improvement over the state-of-the-art (SOTA) for sentiment (SN), while sarcasm (8.34%) and offensiveness (3.58%) saw only moderate gains. This result is expected because sarcasm often masks negative sentiment behind positive wording. We also included a detailed error analysis in Section 5. The error analysis revealed persistent over-reliance on surface features. The result further indicated that the absence of explicit phrases or visual cues in the case of irony and sarcasm leads to misclassification. Besides, the evaluation metrics penalize near-miss predictions, tempering sarcasm and offensiveness improvements. Nevertheless, we find that higher F1 scores align with CoT-based understandings, with Qwen achieving the highest BERT Scores under CoT prompting and LlaVA with ZSC prompting attained an average F1 of 46.12%, representing an 11.2% increase over Exp 1 (QW-FSC: 41.48%) and 25.0% over Exp 2 (QW-LoRA: 36.89%). These results demonstrate that VLM explanation–based fine-tuning outperforms both direct prompting and adapter tuning, and that three-step CoT prompting (ZSC, FSC) produces more structured explanations that enhance downstream LLM fine-tuning, collectively driving significant classification accuracy gains.

5 Error analysis

The classification approach does not fully underline the capability of the VLMs for meme classification. Hence, we further analyze the performance of the best models and prompting techniques using qualitative and quantitative methods.

		Memotion															MAMI						
MDL	PM	HM			SR			OF			SN			MV			MG			MGT			Avg.
		BR	RB	XL	BR	RB	XL	BR	RB	XL	BR	RB	XL	BR	RB	XL	BR	RB	XL	BR	RB	XL	
LV	ZS	26.57	16.24	28.95	26.23	15.97	19.28	26.36	20.52	22.16	26.53	25.56	31.79	51.96	16.75	54.03	66.95	67.80	67.89	53.07	43.29	44.96	35.85
	ZSC	32.01	31.37	29.37	32.70	34.93	**35.59**	35.04	31.97	32.14	**48.22**	44.47	47.87	**58.67**	46.75	55.20	68.52	66.51	68.51	54.33	51.97	57.10	46.12
	FS	35.90	**32.95**	34.59	30.54	34.93	33.01	34.10	30.26	34.35	47.01	43.18	47.32	54.68	53.07	56.75	56.94	60.14	47.68	52.82	48.53	50.52	43.77
	FSC	36.59	31.15	32.85	**34.94**	**35.08**	35.45	31.42	30.53	32.16	47.08	45.39	43.83	48.20	46.75	49.98	61.08	60.09	55.01	50.01	52.07	51.08	39.16
QW	ZS	27.42	11.37	21.47	30.41	18.56	25.24	29.17	21.56	25.46	36.02	25.39	27.32	55.07	37.89	46.58	71.70	67.09	71.38	52.84	56.12	56.56	38.79
	ZSC	31.66	32.44	**38.46**	24.53	26.67	35.33	32.98	28.94	24.52	43.66	**47.51**	**48.31**	57.02	36.75	55.29	71.59	**75.08**	**79.45**	**58.47**	**56.33**	57.56	45.84
	FS	**38.29**	31.43	35.92	32.26	32.02	31.81	**35.61**	32.53	**34.77**	46.57	44.47	39.61	55.01	53.21	**56.95**	33.51	36.43	33.52	56.69	53.83	56.99	41.50
	FSC	35.13	31.89	33.46	33.89	31.29	32.52	34.76	30.64	31.48	41.58	42.17	40.19	55.63	**56.05**	56.49	64.85	68.71	71.46	51.62	54.02	**57.83**	45.51
LM	ZS	29.02	21.37	30.41	28.95	17.78	25.67	28.64	30.35	28.24	37.15	25.39	33.61	46.37	36.75	48.24	**71.73**	71.50	76.01	55.38	56.01	56.06	40.70
	ZSC	30.18	31.37	32.45	28.93	28.91	29.25	34.77	**33.20**	26.97	45.63	35.23	45.59	56.69	53.21	53.97	68.26	70.85	70.40	56.23	56.12	56.33	44.98
IB	ZS	25.15	16.24	21.47	24.62	18.94	29.34	23.20	22.20	17.87	29.63	30.78	25.42	49.35	36.75	42.90	60.15	62.81	60.86	52.34	53.82	52.10	36.00
	ZSC	32.25	30.10	28.37	31.02	28.89	30.03	26.69	28.98	29.28	43.46	33.83	35.39	55.21	55.93	38.51	64.39	63.09	64.81	54.44	54.38	55.65	42.13
Avg.		31.68	26.49	30.65	31.17	27.00	30.21	31.06	28.47	28.28	41.06	36.95	38.85	53.66	44.16	51.24	63.31	64.18	63.92	54.02	53.06	54.69	
Std.		±3.95	±6.28	±3.20	±3.79	±7.13	±5.82	±4.68	±5.49	±6.25	±7.84	±7.29	±7.68	±4.56	±7.35	±6.01	±5.18	±5.84	±6.97	±2.79	±3.27	±2.94	
SOTA			49.09			32.85			34.38			38.30			59.28			87.40			73.14		

Table 5: Finetuning LLMs on VLM-generated explanations (Exp 3). Notation: The highest F1 score of each LLM is in **bold** (column-wise) and the highest F1 score in each task is in blue. LM and IB do not support FS, and FSC prompting.

5.1 Quantitative analysis

We evaluate the textual meme understanding generated by the VLMs in Experiment-3 using standard metrics—BLEU (Post, 2018), ROUGE (Lin, 2004), and BERTScore (Zhang et al., 2019)—following best practices in text generation evaluation (Van der Lee et al., 2021). ROUGE score captures surface-level similarity through n-gram and longest common subsequence overlap. BLEU measures n-gram precision for evaluating fluency and alignment. BERTScore assesses semantic similarity using contextual embeddings, enabling meaning insight beyond lexical matches. The metrics provide complementary insights, including surface-level and semantic understanding required for proper evaluation.

We generated silver-standard textual explanations using GPT-4 (gpt-4-0613) (Achiam et al., 2023) for analyzing the VLM-generated understanding in detail. The datasets do not contain ground-truth explanations. Since generating reliable ground truth through manual annotation in a crowdsourcing framework is a subjective, expensive, resource-intensive, and time-consuming (Liu et al., 2023b; Maniparambil et al., 2023; Xu et al., 2024b), we relied on GPT-4 (gpt-4-0613). Of late, the AI/NLP community is increasingly using LLMs and VLMs for silver-standard annotations (Liu et al., 2023b; Maniparambil et al., 2023; Xu et al., 2024b) where even strong benchmarks now evaluate VLMs using GPT-4 generated references (Liu et al., 2023a; Caffagni et al., 2024; Bhattacharyya and Wang, 2025). Recent computational meme understanding research shows VLMs effectively reduce annotation effort while maintaining quality (Nguyen and Ng, 2024a). Besides, some recent studies have reported that some VLMs can surpass human performance in certain metrics of visual recognition tasks (Lin et al., 2024).

We compare the meme explanations generated by VLMs with those from GPT-4 using the metrics above. Table 6 presents the results of this comparison. We observe a correlation between better textual understanding and improved F1 scores in Table 5. Notably, the QWEN-2-VL model, which performs best in most Experiment-3 cases, also achieves the highest BERTScores across both datasets, indicating stronger semantic understanding.

We analyze the overall quality of the understanding generated by the four pre-trained VLMs under the CoT prompting in Experiment 3. For this purpose, we compare the quality of explanations across correctly and incorrectly classified samples. In order to quantitatively understand the rationale behind the performance of the classifications in Experiment 3. For this purpose, we computed the BERTScore difference between the two groups using GPT-4-generated silver labels as a reference. Figure 2 shows the resulting differences across all tasks and models.

We chose the three-step CoT setting for this analysis because it yielded the highest average F1 score (46.12) in the CoVExFiL experiment. By analyzing CoT-based outputs from the four VLMs, we aim to determine whether larger BERTScore differences reflect meaningful explanations for correct predictions—are associated with better classification performance.

Model	Prompting	Memotion					MAMI				
		ROUGE-1	ROUGE-2	ROUGE-L	BLEU	BERT-Score	ROUGE-1	ROUGE-2	ROUGE-L	BLEU	BERT-Score
LV	ZS	34.75	11.99	22.04	2.23	58.25	34.43	11.52	21.59	2.22	57.88
	ZSC	50.97	20.99	26.58	13.43	64.23	**48.91**	**19.14**	**25.77**	10.89	62.59
	FS	46.75	16.02	24.38	8.41	60.56	38.87	10.20	20.73	5.31	57.02
	FSC	41.44	14.61	23.07	6.76	61.97	37.26	11.32	20.48	5.31	59.75
QW	ZS	33.75	13.69	21.64	3.43	56.18	42.67	15.93	24.09	6.41	60.82
	ZSC	50.99	20.92	26.75	13.32	64.48	41.71	9.52	19.43	4.84	57.36
	FS	43.4	19.2	26.43	5.81	60.01	47.73	17.58	25.2	11.2	**62.78**
	FSC	49.78	19.71	27.45	12.24	**64.96**	47.53	17.49	25.07	11.07	62.65
LM	ZS	44.44	17.65	25.24	8.69	61.57	47.47	17.98	25.43	10.71	61.34
	ZSC	**51.25**	**21.23**	**26.84**	**13.67**	64.57	48.00	18.86	24.32	**11.24**	61.69
IB	ZS	7.17	5.22	6.95	1.1	40.33	7.34	5.08	7.01	0	40.85
	ZSC	7.73	5.31	7.09	1.23	40.34	7.33	5.10	7.01	0	40.81

Table 6: Evaluation performance of the considered VLMs versus GPT-4 generated understanding. The highest value of each metric is in **bold** (column-wise).

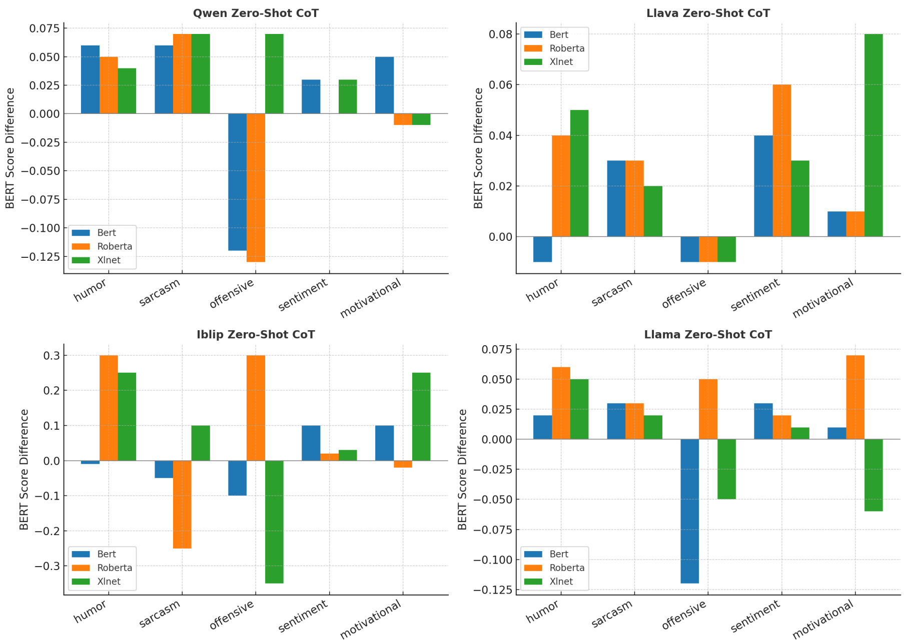


Figure 2: BERT Score Difference between correctly classified vs misclassified samples

To be more specific, higher BERTScore differences indicate that the correctly classified samples are semantically more meaningful. Most likely, these richer explanations guide the downstream LLMs in obtaining accurate predictions during the fine-tuning stage. Conversely, smaller differences suggest that the explanations lacked sufficient semantic depth or clarity. Most likely, the lack of proper comprehension caused misclassification. In such cases, the explanations for misclassified samples often appear to be lexically or semantically similar to the correct predictions. This observation highlights that the model is not able to comprehend subtle linguistic and contextual cues. We came to the following conclusions from Figure 2, indicating the BERT score difference for Memotion datasets.

In some cases, the explanations for incorrect classifications were often lexically similar to those of correct classifications. This contradiction suggests that models are unable to distinguish subtle linguistic and contextual cues in sarcasm and offensiveness. Generally, sarcasm and offensiveness resulted in the lowest BERT Score differences, suggesting that the explanation quality for correct and incorrect classifications is similar. We can conclude that the AI model cannot comprehend the inherent, indirect, and twisted messaging of sarcastic and offensive statements. In contrast, sentiment and humor understanding resulted in relatively higher BERTScore differences, indicating that the VLMs produce a moderate BERT score for the memes they correctly classify. On the contrary, the model fails to comprehend the challenging memes, producing low BERT scores, leading to misclassification. We conclude from the observations that the VLMs are better at identifying emotionally charged language or clear comedic elements in memes. However, motivational classification exhibited mixed performance, with some models relying on generic inspirational language without capturing deeper contextual meaning. Overall, the results highlight that while VLMs can effectively classify emotionally and socially sensitive aspects such as sentiment, they struggle with complex linguistic nuances like sarcasm and offensiveness.

5.2 Qualitative analysis

In this section, we examined misclassified samples, even by the best-performing VLMs in experiment 3. Figure 6 shows a few such examples from the test set. The following are our findings:

As per detailed analysis, we observed several misclassification patterns for diverse meme understanding tasks. These observations revealed the inherent ability of the models to interpret visual and textual cues effectively.

In Example 1, a meme labeled offensive was misclassified as not_offensive. The VLM's explanation focused on childlike expressions and formal attire, ignoring hateful undertones. The highlighted regions were unrelated to offensive cues, revealing the model's inability to detect subtle hate signals.In Example 2, a meme labeled "very_twisted" was predicted as "not_sarcastic." The meme features a Nazi rally with one man refusing to salute, while Hitler is visible in the crowd. The caption reads, "Man is refusing to stand with the fascists that have taken over Germany." Despite the obvious irony, the model failed to detect sarcasm. This failure is likely due to the lack of an explicit mention of Hitler, which resulted in weak alignment between the visual context and the text. Example 3 contained a Big Bang Theory character humorously comparing his singing to Justin Bieber's. The Big Bang Theory character humorously compared his singing to Justin Bieber's. Although the model was able to understand that the meme is humorous, it could not comprehend the severity of the humor as it lacked contextual insights. Consequently, the model misclassified the meme as funny, instead of labelling it as very funny. In Example 4, the model correctly predicted the meme as positive, with the explanation aligning well across both modalities. The model was able to understand the image of a smiling person recalling a funny pet story. The examples demonstrate that the model can identify when visual and textual information are aligned.

As per our analysis, the current Vision-Language Models (VLMs) can understand basic aspects of memes but often miss deeper, context-driven cues. These models frequently rely too heavily on either the visual or textual modality, which causes them to overlook implicit offensiveness, miss sarcasm conveyed through visual irony, or misjudge the degree of humor. They also struggle to recognize culturally or historically significant visual elements when the accompanying text does not make them explicit. Even when the models partially grasp the meme's intent, strict evaluation metrics penalize them harshly, failing to account for close predictions. These findings highlight the need to develop models that align visual and textual information more effectively, reason contextually, and offer clearer explanations to handle subtle and complex tasks in meme understanding.

6 Conclusion

In this treatise, we presented a systematic, in-depth study on utilizing VLMs in diverse strategies for accurate meme classification. The analysis affirmed that pre-trained VLMs generally perform well in the presence of explicit cues in tasks like SN and MG. However, this approach cannot accurately comprehend nuanced content like SR or OF. The prompting strategies, particularly CoT, improved the reasoning and classification accuracy. However, LoRA-based fine-tuning proved to be less effective, mostly because LoRA alters a relatively small number of parameters from very few selective layers. Our proposed CoVExFiL approach, which integrates VLMs and LLMs, performs well in multiple tasks. These findings underscore the effectiveness of prompting and distillation for improving meme understanding.

Our analysis shows that VLMs grasp meme context well when the content contains clear, straightforward clues. However, they struggle with hidden meanings, particularly in OF and SR tasks. The CoT provides moderate performance gains and helps bridge this gap to some extent.

Limitations

The primary limitations of our work are described below: First, we used only publicly available VLMs and did not include larger or closed-source models that might deliver stronger reasoning and different insights. Second, we explored four distinct prompting strategies, yet we did not cover the full spectrum of possible prompt variations, which may affect performance. Third, we used LoRA-based fine-tuning instead of complete model fine-tuning due to high computational costs, possibly limiting adaptability and effectiveness. Fourth, we treated each meme independently and relied on publicly available meme explanations, which may lack external context, such as historical events or social media trends, and reflect annotator biases or miss cultural nuances. Addressing this requires more straightforward guidelines, culturally aligned annotators, and familiarity-based filtering. Addressing these issues through broader model selection, richer evaluation, full fine-tuning, external knowledge integration, and multilingual support offers a clear path for future work.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Tariq Habib Afridi, Aftab Alam, Muhammad Numan Khan, Jawad Khan, and Young-Koo Lee. 2021. A multimodal memes classification: A survey and open research issues. In *Innovations in Smart Cities Applications Volume 4: The Proceedings of the 5th International Conference on Smart City Applications*, pages 1451–1466. Springer.

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katie Millicah, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob Menick, Sebastian Borgeaud, Andrew Brock, Aida Nematzadeh, Sahand Sharifzadeh, Mikolaj Binkowski, Ricardo Barreira, Oriol Vinyals, Andrew Zisserman, and Karen Simonyan. 2022. Flamingo: a visual language model for few-shot learning. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, Red Hook, NY, USA. Curran Associates Inc.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. 2024. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*.

Francesco Barbieri, Jose Camacho-Collados, Luis Espinosa Anke, and Leonardo Neves. 2020. TweetEval: Unified benchmark and comparative evaluation for tweet classification. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1644–1650, Online. Association for Computational Linguistics.

Prithvi Bhattacharya. 2019. Social degeneration through social media: A study of the adverse impact of 'memes'. *2019 Sixth HCT Information Technology Trends (ITT)*, pages 44–46.

Sree Bhattacharyya and James Z. Wang. 2025. Evaluating vision-language models for emotion recognition. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 1798–1820, Albuquerque, New Mexico. Association for Computational Linguistics.

Efrat Blaier, Itzik Malkiel, and Lior Wolf. 2021. Caption enriched samples for improving hateful memes detection. *arXiv preprint arXiv:2109.10649*.

Davide Caffagni, Federico Cocchi, Luca Barsellotti, Nicholas Moratelli, Sara Sarto, Lorenzo Baraldi, Lorenzo Baraldi, Marcella Cornia, and Rita Cucchiara. 2024. The revolution of multimodal large language models: A survey. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 13590–13618, Bangkok, Thailand. Association for Computational Linguistics.

Mu Cai, Zeyi Huang, Yuheng Li, Utkarsh Ojha, Haohan Wang, and Yong Jae Lee. 2025. An investigation on llms' visual understanding ability using svg for image-text bridging. In *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 5377–5386. IEEE.

Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, Kaijie Zhu, Hao Chen, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, Wei Ye, Yue Zhang, Yi Chang, Philip S. Yu, Qiang Yang, and Xing Xie. 2024. A survey on evaluation of large language models. *ACM Trans. Intell. Syst. Technol.*, 15(3).

Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. 2020. Uniter: Universal image-text representation learning. In *European conference on computer vision*, pages 104–120. Springer.

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. 2023. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality.

Wenliang Dai, Junnan Li, DONGXU LI, Anthony Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale N Fung, and Steven Hoi. 2023. Instructblip: Towards general-purpose vision-language models with instruction tuning. In *Advances in Neural Information Processing Systems*, volume 36, pages 49250–49267. Curran Associates, Inc.

Shiv Shankar Das. 2023. Rising popularity of internet memes in india: A media psychology perspective. *Indian Journal of Social Science and Literature*, 2(3):5–9.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pages 4171–4186.

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.

Katharine Gelber and Luke McNamara. 2016. Evidencing the harms of hate speech. *Social Identities*, 22(3):324–341.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Kapil Gupta and Cheshta Kwatra. 2021. Multimodal meme sentiment analysis with image inpainting. In *2021 IEEE 8th Uttar Pradesh Section International Conference on Electrical, Electronics and Computer Engineering (UPCON)*, pages 1–6. IEEE.

Ming Shan Hee, Shivam Sharma, Rui Cao, Palash Nandi, Tanmoy Chakraborty, and Roy Ka-Wei Lee. 2024. Recent advances in hate speech moderation: Multimodality and the role of large models. *arXiv preprint arXiv:2401.16727*.

Lisa Anne Hendricks, Ronghang Hu, Trevor Darrell, and Zeynep Akata. 2018. Grounding visual explanations. In *Proceedings of the European conference on computer vision (ECCV)*, pages 264–279.

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.

Prince Jha, Raghav Jain, Konika Mandal, Aman Chadha, Sriparna Saha, and Pushpak Bhattacharyya. 2024. Memeguard: An llm and vlm-based framework for advancing content moderation via meme intervention. *arXiv preprint arXiv:2406.05344*.

Robin Jia and Percy Liang. 2017. Adversarial examples for evaluating reading comprehension systems. *arXiv preprint arXiv:1707.07328*.

Kanika Kalra, Bhargav Kurma, Silpa Vadakkeeveetil Sreelatha, Manasi Patwardhan, and Shirish Karande. 2020. Understanding advertisements with bert. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7542–7547.

Gitanjali Kumari, Dibyanayan Bandyopadhyay, Asif Ekbal, Arindam Chatterjee, and Vinutha B.N. 2025. Identifying offensive memes in low-resource languages: A multi-modal multi-task approach using valence and arousal. *Computer Speech Language*, 92:101781.

Gitanjali Kumari, Dibyanayan Bandyopadhyay, Asif Ekbal, Santanu Pal, Arindam Chatterjee, and Vinutha BN. 2024. Let's all laugh together: A novel multitask framework for humor detection in internet memes. *IEEE Transactions on Computational Social Systems*, 11(3):4385–4395.

Gwang Gook Lee and Mingwei Shen. 2022. Multimodal, multi-task learning for memotion 2.0 challenge.

Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. 2019. Visualbert: A simple and performant baseline for vision and language. *arXiv preprint arXiv:1908.03557*.

Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, et al. 2020. Oscar: Object-semantics aligned pre-training for vision-language tasks. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXX 16*, pages 121–137. Springer.

Zongxia Li, Xiyang Wu, Hongyang Du, Fuxiao Liu, Huy Nghiem, and Guangyao Shi. 2025. A survey of state of the art large vision language models: Benchmark evaluations and challenges. In *Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR) Workshops*, pages 1587–1606.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

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*, WWW '24, page 2359–2370, New York, NY, USA. Association for Computing Machinery.

Chen Liu, Gregor Geigle, Robin Krebs, and Iryna Gurevych. 2022. Figmemes: A dataset for figurative language identification in politically-opinionated memes. In *Proceedings of the 2022 conference on empirical methods in natural language processing*, pages 7069–7086.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2024a. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26296–26306.

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

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023a. Visual instruction tuning. In *Advances in Neural Information Processing Systems*, volume 36, pages 34892–34916. Curran Associates, Inc.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023b. G-eval: Nlg evaluation using gpt-4 with better human alignment. *arXiv preprint arXiv:2303.16634*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *Advances in neural information processing systems*, 32.

Mayug Maniparambil, Chris Vorster, Derek Molloy, Noel Murphy, Kevin McGuinness, and Noel E O'Connor. 2023. Enhancing clip with gpt-4: Harnessing visual descriptions as prompts. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 262–271.

Delfina S. Martinez Pandiani, Erik Tjong Kim Sang, and Davide Ceolin. 2025. 'toxic' memes: A survey of computational perspectives on the detection and explanation of meme toxicities. *Online Social Networks and Media*, 47:100317.

Khoi Nguyen and Vincent Ng. 2024a. Computational meme understanding: A survey. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 21251–21267.

Khoi P. N. Nguyen and Vincent Ng. 2024b. Computational meme understanding: A survey. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 21251–21267, Miami, Florida, USA. Association for Computational Linguistics.

Sofiane Ouaari, Tsegaye Misikir Tashu, and Tomáš Horváth. 2022. Multimodal feature extraction for memes sentiment classification. In *2022 IEEE 2nd Conference on Information Technology and Data Science (CITDS)*, pages 285–290. IEEE.

Hongliang Pan, Zheng Lin, Peng Fu, Yatao Qi, and Weiping Wang. 2020. Modeling intra and inter-modality incongruity for multi-modal sarcasm detection. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1383–1392.

Delfina S Martinez Pandiani, Erik Tjong Kim Sang, and Davide Ceolin. 2025. 'toxic'memes: A survey of computational perspectives on the detection and explanation of meme toxicities. *Online Social Networks and Media*, 47:100317.

Yulia Petrova. 2021. Meme language, its impact on digital culture and collective thinking. In *E3S Web of Conferences*, volume 273, page 11026. EDP Sciences.

Matt Post. 2018. A call for clarity in reporting bleu scores. *arXiv preprint arXiv:1804.08771*.

Shraman Pramanick, Shivam Sharma, Dimitar Dimitrov, Md Shad Akhtar, Preslav Nakov, and Tanmoy Chakraborty. 2021. Momenta: A multimodal framework for detecting harmful memes and their targets. *arXiv preprint arXiv:2109.05184*.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR.

Sathyanarayanan Ramamoorthy, Nethra Gunti, Shreyash Mishra, S Suryavardan, Aishwarya Reganti, Parth Patwa, Amitava DaS, Tanmoy Chakraborty, Amit Sheth, Asif Ekbal, et al. 2022. Memotion 2: Dataset on sentiment and emotion analysis of memes. In *Proceedings of De-Factify: Workshop on Multimodal Fact Checking and Hate Speech Detection, CEUR*.

Siddhant Bikram Shah, Shuvam Shiwakoti, Maheep Chaudhary, and Haohan Wang. 2024. Memeclip: Leveraging clip representations for multimodal meme classification. *arXiv preprint arXiv:2409.14703*.

Chhavi Sharma, Deepesh Bhageria, William Scott, Srinivas Pykl, Amitava Das, Tanmoy Chakraborty, Viswanath Pulabaigari, and Bjorn Gamback. 2020. Semeval-2020 task 8: Memotion analysis–the visuo-lingual metaphor! *arXiv preprint arXiv:2008.03781*.

Shivam Sharma, Siddhant Agarwal, Tharun Suresh, Preslav Nakov, Md Shad Akhtar, and Tanmoy Chakraborty. 2023. What do you meme? generating explanations for visual semantic role labelling in memes. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 9763–9771.

Shivam Sharma, Md Shad Akhtar, Preslav Nakov, and Tanmoy Chakraborty. Disarm: Detecting the victims targeted by harmful memes.

Shivam Sharma, Firoj Alam, Md. Shad Akhtar, Dimitar Dimitrov, Giovanni Da San Martino, Hamed Firooz, Alon Halevy, Fabrizio Silvestri, Preslav Nakov, and Tanmoy Chakraborty. 2022. Detecting and understanding harmful memes: A survey. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI-22*, pages 5597–5606. International Joint Conferences on Artificial Intelligence Organization. Survey Track.

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.

Tiberiu Sosea and Cornelia Caragea. 2025. Hard emotion test evaluation sets for language models. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 7930–7944, Albuquerque, New Mexico. Association for Computational Linguistics.

Tipu Sultan, Mohammad Abu Tareq Rony, Mohammad Shariful Islam, Saad Aldosary, and Walid El-Shafai. 2024. Memesvita: A novel multimodal fusion technique for troll memes identification. *IEEE Access*.

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.

Chris Van der Lee, Albert Gatt, Emiel Van Miltenburg, and Emiel Krahmer. 2021. Human evaluation of automatically generated text: Current trends and best practice guidelines. *Computer Speech & Language*, 67:101151.

Danae Sánchez Villegas, Saeid Mokaram, and Nikolaos Aletras. 2021. Analyzing online political advertisements. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3669–3680.

Bingbing Wang, Shijue Huang, Bin Liang, Geng Tu, Min Yang, and Ruifeng Xu. 2024. What do they "meme"? a metaphor-aware multi-modal multi-task framework for fine-grained meme understanding. *Knowledge-Based Systems*, 294:111778.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Jialu Xing, Jianping Liu, Jian Wang, Lulu Sun, Xi Chen, Xunxun Gu, and Yingfei Wang. 2024. A survey of efficient fine-tuning methods for vision-language models—prompt and adapter. *Computers & Graphics*, 119:103885.

Bo Xu, Tingting Li, Junzhe Zheng, Mehdi Naseriparsa, Zhehuan Zhao, Hongfei Lin, and Feng Xia. 2022. Met-meme: A multimodal meme dataset rich in metaphors. In *Proceedings of the 45th international ACM SIGIR conference on research and development in information retrieval*, pages 2887–2899.

Yanzhi Xu, Yueying Hua, Shichen Li, and Zhongqing Wang. 2024a. Exploring chain-of-thought for multimodal metaphor detection. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 91–101, Bangkok, Thailand. Association for Computational Linguistics.

Zhiyang Xu, Chao Feng, Rulin Shao, Trevor Ashby, Ying Shen, Di Jin, Yu Cheng, Qifan Wang, and Lifu Huang. 2024b. Vision-flan: Scaling human-labeled tasks in visual instruction tuning. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 15271–15342.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. *Advances in neural information processing systems*, 32.

Francisco Yus. 2019. Multimodality in memes: A cyberpragmatic approach. *Analyzing digital discourse: New insights and future directions*, pages 105–131.

Jing Zhang and Yujin Wang. 2022. Srcb at semeval-2022 task 5: Pretraining based image to text late sequential fusion system for multimodal misogynous meme identification. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, pages 585–596.

Jingyi Zhang, Jiaxing Huang, Sheng Jin, and Shijian Lu. 2024. Vision-language models for vision tasks: A survey. *IEEE Trans. Pattern Anal. Mach. Intell.*, 46(8):5625–5644.

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

Yazhou Zhang, Jinglin Wang, Yaochen Liu, Lu Rong, Qian Zheng, Dawei Song, Prayag Tiwari, and Jing Qin. 2023. A multitask learning model for multimodal sarcasm, sentiment and emotion recognition in conversations. *Information Fusion*, 93:282–301.

Yang Zhong and Bhiman Kumar Baghel. 2024. Multimodal understanding of memes with fair explanations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2007–2017.

Qiji Zhou, Ruochen Zhou, Zike Hu, Panzhong Lu, Siyang Gao, and Yue Zhang. 2024. Image-of-thought prompting for visual reasoning refinement in multimodal large language models. *arXiv preprint arXiv:2405.13872*.

Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. 2023. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*.

A Related Works

In recent years, there has been a growing interest in analyzing memes through both custom deep learning (DL) architectures and vision-language models (VLMs). A comprehensive survey by Nguyen and Ng (2024b) compiled 24 datasets spanning five major task categories related to meme understanding. Complementing this, Martinez Pandiani et al. (2025) examined 158 content-based studies, including 119 newly analyzed papers and identified more than 30 datasets specifically focused on toxic meme analysis published up to 2024.

Over the past few years, multi-modal meme analysis has emerged as a prominent research area in NLP and multi-modal learning(Nguyen and Ng, 2024a; Kumari et al., 2025; Wang et al., 2024; Nguyen and Ng, 2024b; Pandiani et al., 2025; Sharma et al., 2022). These works generally focused on understanding the interplay of visual and textual information in memes for addressing a wide range of tasks, such as sentiment and emotion detection, humor and sarcasm recognition, identification of misogynistic or offensive content, figurative language interpretation, and bias assessment in model predictions (Afridi et al., 2021). Several specialized corpora have been released to capture the rich variety of figurative, humorous, and harmful content in memes. Xu et al. (2022) introduced MET-meme, a collection targeted at metaphorical memes, demonstrating that models struggle when literal text masks non-literal intent. Liu et al. (2022) introduced the FigMemes dataset containing memes consisting of sarcasm, irony, hyperbole, and metaphor. The work also demonstrated that integrating social-political context via the MSDBert architecture proposed in the work substantially improves the detection of figurative language. Suryawanshi et al. (2020) proposed the MultiOFF meme dataset, annotated for fine-grained offensiveness and satire, highlighting the role of cultural and linguistic factors in perceiving offense. Fersini et al. (2022) curated the misogyny-focused SemEval-2022 Task 5 benchmark, which goes beyond binary labels to classify different types of gendered hate. Finally, the Memotion and MAMI datasets provide multi-dimensional labels (humor, sentiment, intent), enabling simultaneous evaluation of positive and harmful meme aspects (Ramamoorthy et al., 2022; Afridi et al., 2021).

Researchers have developed sophisticated deep learning architectures that explicitly model both intra-modal and inter-modal interactions, integrating textual and visual information for meme classification. Pan et al. (2020) introduced a dual attention framework with separate intra-modal and inter-modal blocks designed to capture contradictions between text and image, achieving significant improvements in sarcasm detection performance. Similarly, Liu et al. (2022) developed MSD-Bert, which fuses cross-modal embeddings within a BERT backbone to better capture subtle figurative language cues in the FigMemes dataset (Liu et al., 2022). Sharma et al. introduced a framework, termed DISARM (Detecting the Victims Targeted by Harmful Memes), to augment standard multi-modal fusion with Named Entity Recognition. Gupta and Kwatra (Gupta and Kwatra, 2021) combined image inpainting with a Vision Transformer architecture to enhance visual cue recovery for improved meme sentiment analysis. Ouaari et al. (2022) presented a multimodal feature extraction method using deep visual and textual encoders for classifying meme sentiment. Sultan et al. (2024) proposed MemesViTa, a novel hierarchical fusion model based on Vision Transformers for accurate detection of troll memes. Collectively, these studies highlight that sophisticated fusion techniques designed to capture inter-modal dynamics are crucial for achieving substantial gains across diverse meme analysis sub-tasks, including humor, sarcasm, and harm detection. Early works like Hendricks et al. (2018) introduced grounded visual explanations, aligning model justifications with image regions. Jia and Liang (2017) showed that explanation generation enhances trust in NLP models, a principle now extended to multi-modal meme understanding. Radford et al. (2021) demonstrated the ability of CLIP to generate interpretable outputs for memes by aligning visual and textual modalities. Sharma et al. (2023) proposed LUMEN, a framework for humor classification in memes, combining multimodal understanding with contextual explanation generation. Parallel to the direct multimodal analysis of memes, another line of research has applied language models to deconstruct other forms of persuasive media, albeit in a unimodal, text-only context. For instance, Kalra et al. (2020) focused on commercial advertisements by creating a new dataset annotated with fine-grained rhetorical strategies—such as appeals to emotion or scarcity—and then fine-tuned BERT to classify these techniques from ad text. In the domain of political advertising, Villegas et al. (2021) first trained a text-based classifier to identify a political ad's sponsoring party and subsequently performed a differential analysis to reveal the distinct topics and persuasive frames characteristic of each party's messaging. These studies underscore the value of using transformer models to decode complex, persuasive language in specific cultural domains, a principle our work extends to the more challenging multimodal context of memes. VLMs, meanwhile, have revolutionized meme analysis by enabling semantically aligned, multi-modal representations. VisualBERT (Li et al., 2019) and ViLBERT (Lu et al., 2019) introduced unified and dual-stream architectures, respectively, for vision-language fusion. UNITER (Chen et al., 2020) and OSCAR (Li et al., 2020) enhanced alignment through large-scale pretraining and object tag anchoring. CLIP (Radford et al., 2021) set a new standard with contrastive learning and zero-shot classification. Models like MiniGPT-4 (Zhu et al., 2023) extended these capabilities by incorporating visual grounding and transformer-based vision processing. Additional strategies, such as caption enrichment (Blaier et al., 2021), multi-task learning (Lee and Shen, 2022), and external knowledge integration (Pramanick et al., 2021), have been shown to improve meme classification performance and interpretability. (Jha et al., 2024) proposed a MemeGuard framework fine-tuned with a dedicated VLM for harmful meme interpretation, applied a multi-modal knowledge-selection module, and then prompts a general-purpose LLM to generate context-aware interventions.

Despite advancements in meme understanding, current deep learning models often fail to generalize beyond the datasets they are trained on, limiting their real-world applicability. Most lack **interpretability**, offering no insight into their predictions—an issue in sensitive contexts. Additionally, **robustness** to adversarial content and **generalization** across cultures remain largely unaddressed. Addressing these gaps is key to building adaptable, trustworthy meme understanding systems.

B Experimental Setup

B.1 Datasets Details

This section includes the details of the two datasets, namely, **Memotion** dataset Sharma et al. (2020) and Multimedia Automatic Misogyny Identification dataset (**MAMI** hereafter) Fersini et al. (2022), used for extensive experimentation. The dataset statistics are given in Table 1, and dataset characteristics are detailed in the following subsection-

B.1.1 Memotion Dataset

The memotion dataset (Sharma et al. (2020)) (publicly available on Kaggle) contains 9,871 multimodal memes collected from Google images. The dataset creation process involved selecting memes from 52 categories, including political figures and popular cultural references such as Hillary, Trump, Minions, and Baby Godfather. Only memes with embedded English text were retained to ensure linguistic consistency. The memes were annotated using Amazon Mechanical Turk[1], where five annotators independently judged each meme, and a majority voting scheme was used to combine the annotations. The memes were labeled according to five emotion-oriented dimensions: *humor, sarcasm, offensiveness, motivation, and sentiment*. In particular, the details of the classification tasks are as follows-

- **Task A: Sentiment Classification**
  Classifies memes into three classes: *Positive*, *Neutral*, and *Negative*.
- **Task B: Emotion Classification**
  Labels memes based on one of the following emotions: *humor*, *Sarcasm*, *Offensiveness*, or *Motivation*.
- **Task C: Scales of Semantic Classes**
  Categories memes into overlapping emotional scales:
  - **Humor**: not_funny, funny, very_funny, hilarious.
  - **Sarcasm**: not_sarcastic, general, twisted_meaning, very_twisted.
  - **Offensiveness**: not_offensive, slight, very_offensive, hateful_offensive.
  - **Motivation**: not_motivational, motivational.

We have visualized some sample memes and their corresponding class labels from the Memotion dataset in Fig. 2.

[1] https://www.mturk.com/

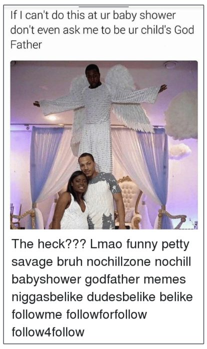
Image	Tasks				
	humor	**sarcasm**	**offensive**	**sentiment**	**motivational**
	‣ hilarious ‣ **very_funny** ‣ funny ‣ not_funny	‣ very_twisted ‣ **twisted_meaning** ‣ general ‣ not_sarcastic	‣ hateful_offensive ‣ very_offensive ‣ slight ‣ **not_offensive**	‣ postive ‣ **neutral** ‣ negative	‣ **motivational** ‣ not_motivational
	‣ hilarious ‣ very_funny ‣ **funny** ‣ not_funny	‣ very_twisted ‣ twisted_meaning ‣ **general** ‣ not_sarcastic	‣ hateful_offensive ‣ very_offensive ‣ **slight** ‣ not_offensive	‣ postive ‣ neutral ‣ **negative**	‣ **motivational** ‣ not_motivational

Figure 3: Data samples from the Memotion dataset. For each meme, the full set of subcategories corresponding to each classification task is listed. The ground-truth label for each task is highlighted in green.

B.1.2 MAMI

The MAMI dataset Fersini et al. (2022) (Apache 2.0 licensed) detects misogyny in memes and subclassifies the type of misogynist content. The memes in this dataset were collected from two sources- social media platforms such as Twitter and Reddit, and meme-sharing websites like 9GAG, Imgur, and KnowYourMeme. These memes were collected by using specific hashtags such as #girl, #girlfriend, #women, #feminist, threads, and discussions covering feminist debates and other similar events. The dataset was annotated using the crowdsourcing concept. The dataset consists of two types of labels: i) whether the meme is misogynist or not, and ii) whether the misogynist meme corresponds to shaming, stereotype, objectification, or violence classes. The first stage involved assigning binary labels to detect the presence or absence of misogyny and subclassification of misogynous memes. Figure 4 presents a few sample memes from the MAMI dataset for manual understanding of its content and labeling structure. The dataset supports two key tasks:

The details of the sub-tasks is given below-

- **Sub-task A: Misogyny Detection**
  Classifies memes as *Misogynous* or *Non-Misogynous*.

- **Sub-task B: Misogyny Type Classification**
  Categorizes misogynous memes into one or more overlapping types:
  - **shaming**: Criticism of women based on appearance or behavior.
  - **Stereotype**: Imposition of traditional roles or fixed traits on women.
  - **Objectification**: Treating women as objects.
  - **violence**: Depictions or implications of physical or psychological violence against women.

Table 1 summarizes the sample distribution for both the Memotion and the MAMI datasets. "Avg. Length" denotes the average sample length in words for each category.

B.2 Configuration of VLMs

In this study, we evaluate four open-source vision-language models (VLMs) for the meme classification task, as they are widely used in the image captioning and visual reasoning literature. The specific models used in our experiments are listed in Table 2, along with their implementation details.

LLAVA-1.6: LLAVA-1.6 (Large Language and Vision Assistant) (LV) (Liu et al., 2024b) is a vision-language model optimized for multi-modal tasks. It extends LLAVA-1.5 (Liu et al., 2024a) by incorporating improved instruction tuning and

Image	Task A	Task B			
	misogyny	**shaming**	**stereotype**	**objectification**	**violence**
	**1**	**0**	**1**	**0**	**1**
	**0**	**0**	**0**	**0**	**0**

Figure 4: Example from the MAMI dataset illustrating two tasks: Task A (binary: 1 = misogynistic, 0 = non-misogynistic) and Task B (multi-label: 1/0 for shaming, stereotype, objectification, violence).

enhanced vision-language alignment, where CLIP (Radford et al., 2021) is used as the vision encoder and Vicuna (Chiang et al., 2023) as the text encoder.

Qwen2-VL: Qwen2-VL (Bai et al., 2024) is an advanced vision-language model, specifically designed to employ a mixed training regimen, both image and text.The model integrates a Vision Transformer (ViT) for visual processing with a Qwen2 series language model for textual understanding. A standout feature is its use of Multimodal Rotary Position Embedding (M-RoPE), which effectively encodes positional information across modalities.

LLaMA-3.2-Vision: Building on the foundation of LLaMA-3.1 (Grattafiori et al., 2024), LLaMA-3.2-Vision incorporates a 40-layer transformer enhanced with a dedicated vision adapter. This adapter leverages local and global cross-attention mechanisms to effectively merge image features from a vision transformer into the pre-trained language model. The model's scalable architecture is designed to handle sophisticated multimodal reasoning and process large-scale datasets, making it suitable for the challenges of meme interpretation.

InstructBLIP: InstructBLIP (Dai et al., 2023) is a Vision-Language Model (VLM) fine-tuned on a wide array of vision-language instructions. As an evolution of BLIP-2, it employs a Q-Former component to bridge a frozen vision encoder with a powerful language model, such as Vicuna or Flan-T5. Its training on diverse instructional data equips the model with strong zero-shot generalization capabilities, allowing it to perform proficiently on various tasks, including meme classification, without prior task-specific training.

B.2.1 Configuration of LLMs

To complement the vision-language understanding from VLMs, we use three pretrained language models—BERT (BR) (Devlin et al., 2019), RoBERTa (RB) (Liu et al., 2019), and XLNet (XL) (Yang et al., 2019) —as classification backbones in our CoVExFiL pipeline (Exp 3). We have listed the particulars of the LLMs in Table 7. These models are fine-tuned using the textual explanations generated by the VLMs, enabling a decoupled two-step classification process.

BERT (BR): BERT (Bidirectional Encoder Representations from Transformers) (Devlin et al., 2019) is a transformer-based model trained using masked language modeling. Its bidirectional attention makes it effective for capturing contextual dependencies in text. We use the base uncased version in our experiments.

RoBERTa (RB): RoBERTa (Liu et al., 2019) improves upon BERT by training on larger datasets with dynamic masking and no next-sentence prediction. It is known for robust performance across many classification benchmarks.

XLNet (XL): XLNet (Yang et al., 2019) is an autoregressive model that incorporates permutation-based training to capture the bidirectional context. It overcomes some of the limitations of BERT, particularly in modeling word order and long-range dependencies.

Prompt Type	Prompt template
Zero-shot	**{Image:** {image}, **Query** Analyze the following meme and classify its category as one of the following if the given meme is: [class1, class2, class3, class4]. Also, provide a brief explanation for your classification" **}** **Answer:**

Few-shot (k=1)	**Support Examples:**{
	[image1] [text description for class 1] [Label: 1]
	[image2] [[text description for class 2] [Label: 2]
	[image3] [[text description for class class 3] [Label: 3]
	[image4] [[text description for class class 4] [Label: 4]
	}
	{Image: {image},
	Query: Analyze the following meme and classify its category as one of the following if the given meme is: [class1, class2, class3, class4]. Also, provide a brief explanation for your classification"**}**
	Answer:

Chain-of-thought	**{Image:** {image}, **Query:**"Analyze the following meme and classify its category as one of the following if the given meme is:[class1, class2, class3, class4]. Let's think step by step. Step1: Start by analyzing the visual content to describe what is seen in the image. Step2:Then examine the textual content to explain the meaning or implication of the text. Step3:Finally, provide a combined interpretation of how the visual and textual elements work together to convey a message. Step4:Conclude with the classification based on your reasoning.**}** **Answer:**

Few-shot Chain-of-thought	**Support Examples:**{
	[image1] [text description for class 1] [Label: 1]
	[image2] [[text description for class 2] [Label: 2]
	[image3] [[text description for class class 3] [Label: 3]
	[image4] [[text description for class class 4] [Label: 4]
	}
	{Image: {image},
	Query:"Analyze the following meme and classify its category as one of the following if the given meme is:[class1, class2, class3, class4].
	Let's think step by step.
	Step1: Start by analyzing the visual content to describe what is seen in the image.
	Step2:Then examine the textual content to explain the meaning or implication of the text.
	Step3:Finally, provide a combined interpretation of how the visual and textual elements work together to convey a message.
	Step4:Conclude with the classification based on your reasoning.**}**
	Answer:

Figure 5: Prompt Template for Experiment 1. Here, we have specified the prompts we used in ZS, ZSC, FS, and FSC.

LLM	Size	Version
BERT	110M	`bert-base-uncased`
RoBERTa	125M	`roberta-base`
XLNet	110M	`xlnet-base-cased`

Table 7: Considered LLMs and their corresponding versions for our experiments.

B.2.2 Baselines

The baseline models presented in Table 8 serve as reference points for evaluating performance on the Memotion and MAMI datasets. These models were selected based on their relevance to the respective tasks and their reported effectiveness in prior studies.

- M2Seq2Seq-MLD (Zhang et al., 2023): A multi-task seq2seq model for multimodal sarcasm, sentiment, and emotion recognition. It captures intra- and inter-modality/context-task dynamics for SN and HM.
- MT-BERT+TextGCN (Kumari et al., 2024): Combines Multi-Task BERT with TextGCN for improved text classification via contextual and graph-based features.
- PBR (Pretraining-Based Representation) (Zhang and Wang, 2022): This model combines image features from CLIP with textual embeddings from BERT and UNITER, integrated through a late fusion strategy. The final predictions are made using an ensemble of XGBoost classifiers and refined through rule-based post-processing.
- BERT+ViT (Singh et al., 2023): This model combines a variant of the BERT language model, pretrained on hate-speech text data, with a Vision Transformer (ViT) serving as the visual encoder.

B.3 Configuration of Prompts

In Experiment 1, we explored a range of prompting strategies to evaluate their effectiveness in meme understanding. These included Zero-Shot (ZS), Zero-Shot Chain-of-Thought (ZSC), Few-Shot (FS), and Few-Shot Chain-of-Thought (FSC)

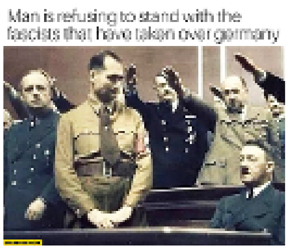
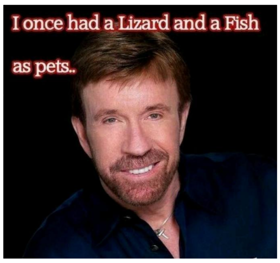
MEME	VLM Meme Understanding CoT
 Label:**hateful_offensive** Predicted:**not_offensive**	Visual Content Analysis The meme consists of two images arranged vertically. 1. Top Image: Features a baby dressed in a formal tuxedo. The baby has a serious expression and is pointing, suggesting authority or a commanding demeanor. The background appears to be an outdoor setting, possibly a walkway or park. 2. Bottom Image: Shows a young boy, also in formal attire (a tuxedo), interacting with a woman in a wedding dress. The boy is looking up at the woman, who appears to be smiling or looking fondly at him. This scene implies a social or affectionate moment were he's teasing her. Textual Content Examination 1. Top Image Text: WOMAN, GET OVER HERE THIS INSTANT - This text is written in a bold, prominent font. The phrase suggests a humorous take on the baby's serious demeanor, as it mimics a commanding or authoritative tone typically used by an adult. 2. Bottom Image Text: Hmm. not bad. - This text is more casual and laid-back, indicating the boy's evaluation or approval of the woman, likely perceiving her attractiveness or presence positively. Combined Interpretation The meme juxtaposes the two images to create a humorous narrative.
 Label:**very_twisted** Predicted:**not_sarcastic**	Visual Content Analysis The image depicts a historical scene involving multiple figures, likely from a Nazi rally or meeting. Central to the composition is a man in a brown uniform with a red armband, standing with his arms crossed and an expression of defiance or discomfort. He is surrounded by several individuals, predominantly wearing military uniforms, who are raising their right arms in a salute, a gesture associated with Nazi ideology. To the right, a recognizable figure appears to be Adolf Hitler, seated and looking on. The overall tone of the image conveys tension, with a clear division between the standing man and the others who are participating in the salute. Textual Content Examination The text accompanying the image states: Man is refusing to stand with the fascists that have taken over Germany. This phrase underscores the central theme of resistance against authoritarianism. It highlights the defiance of the man in the center, suggesting he is standing against the prevailing fascist ideology represented by Hitler and his supporters. The use of the word refusing emphasizes an act of willful opposition, further reinforcing the idea of individual courage in the face of a powerful regime. Combined Interpretation The interplay between the visual and textual elements constructs a narrative of resistance by a man. The image visually
 Label:**very_funny** Predicted:**funny**	This meme features a scene from the television show ""The Big Bang Theory."" The image shows a character holding a microphone, seemingly in the middle of a performance or speech. The text above the character This meme effectively combines sarcasm and reality. reads ""STILL A BETTER SINGER THAN JUSTIN BEIBER,"" which is a humorous comparison suggesting that the character in the image is a better singer than the pop star Justin Bieber. The humor comes from the absurdity of the comparison—given that Justin Bieber is a well-known and popular singer, it is unlikely that someone on a TV show would claim to be a better singer than him. The juxtaposition of a fictional character's claim with a real-world celebrity's reputation creates a funny and exaggerated scenario. The meme is likely intended to be lighthearted and entertaining, playing on the absurdity of the comparison."
 Label:**positive** Predicted:**postive**	The image features a person with short, light brown hair and a mustache. The text overlay on the image reads, I once had a Lizard and a Fish as pets... The text is written in red and white, with the words Lizard and a Fish in white and the rest of the text in red. ### Analysis and Explanation 1. **Visual Content**: - The person appears to be smiling, suggesting a positive or lighthearted tone. - The choice of a black background highlights the person, making them the focal point of the image. 2. **Textual Content**: - The text I once had a Lizard and a Fish as pets... implies that the person had pets that are not typically kept together in the same environment. Lizards and fish are often kept in separate habitats, with fish needing water and lizards needing land. - The text suggests a humorous or unexpected anecdote, possibly indicating that the person had a unique or unusual pet setup. ### Combined Interpretation The combination of the visual and textual elements creates a humorous and light-hearted meme. The person's positive expression and the unexpected nature of having a lizard and a fish as pets together add to the comedic effect. The stark contrast between the black background and the person's clothing further emphasizes the subject, drawing attention to the text and the person's reaction. In conclusion, the meme uses the juxtaposition of a smiling person and a humorous pet story to create a funny and relatable

Figure 6: Examples from the test set with their corresponding gold labels are shown to illustrate the VLM's understanding. Memes where the task was performed well are marked in green, those performed moderately well or relatable are marked in blue, and those where the task was performed poorly are marked in red.

prompting. The specific prompts used in each strategy, along with their corresponding examples and experimental settings, are illustrated in Figure 5. This figure presents detailed query formats and prompt structures for all four strategies, making it easier to understand the differences and design choices involved. Also for reproducibility we kept the temperature of the VLM very low (0.1).

We used string presence through label parsing to determine the predicted class from the model's output. This label parsing strategy was applied consistently across all tasks to extract the appropriate

Category	Labels	SOTA	Model
Sentiment	Pos, Neu, Neg	38.30	M2Seq2Seq-MLD
Humour	[H, VF, F, NF]	49.09	M2Seq2Seq-MLD
Sarcasm	[VT, TM, S, G]	32.85	MT-BERT+TextGCN
Offense	[HF, VF, O, NO]	34.38	MT-BERT+TextGCN
Motivation	[M, NM]	59.28	MT-BERT+TextGCN
Misogyny	[Miso, Non-Miso]	87.40	BERT+ViT
Misogyny Type	[V, S, O, Sh]	73.14	PBR

Table 8: Performance comparison across different tasks and metrics after removing baseline results.

prediction whenever a class name appeared in the response.

B.4 Details of Computational Efficiency

On average, LoRA fine-tuning large VLMs on an A100 80GB GPU required 6-9 GPU hours and 44-60 GB of memory per task, whereas fine-tuning smaller LLMs with CoVExFiL reduced these requirements to 0.2-0.4 GPU hours and 10-12 GB of memory per task. For inference, both the time per sample and GPU memory usage were consistent across the prompting and CoVExFiL methods, as both approaches utilized prompting strategies. The average inference time per sample remained 8-18 seconds, and GPU memory usage stayed at 16-40 GB. Furthermore, as the dataset size increased, both GPU hours and memory usage also increased.