MemeSense: An Adaptive In-Context Framework for Social Commonsense Driven Meme Moderation

⚠DISCLAIMER: This paper features memes that some readers may find vulgar/offensive/hateful.

Anonymous ACL submission

Abstract

Memes present unique moderation challenges due to their subtle, multimodal interplay of images, text, and social context. Standard systems relying predominantly on explicit textual cues often overlook harmful content camouflaged by irony, symbolism, or cultural references. To address this gap, we introduce MemeSense, an adaptive in-context learning framework that fuses social commonsense reasoning with visually and semantically related reference examples. By encoding crucial task information into a learnable cognitive shift vector, MemeSense effectively balances lexical, visual, and ethical considerations, enabling precise yet context-aware meme intervention. Extensive evaluations on a curated set of implicitly harmful memes demonstrate that MemeSense substantially outperforms strong baselines, paving the way for safer online communities. We will make the dataset and source code publicly available upon acceptance.

1 Introduction

Memes have emerged as a powerful form of online expression, where seemingly lighthearted humor can conceal offensive, derogatory, or culturally charged subtexts. Their multimodal nature combining images, text, and symbolism poses significant hurdles for content moderation systems, especially those built primarily around textual analysis (Maity et al., 2024; Jain et al., 2023; Jha et al., 2024b,a). Large vision-language models (VLMs), including GPT-4o (OpenAI et al., 2024), Gemini 2.0 (Team et al., 2024), and Qwen 2.5 (Qwen et al., 2025), often show reduced accuracy on image-centric memes precisely because they depend heavily on overt text clues (Sharma et al., 2023; Agarwal et al., 2024). In contrast, humans effortlessly parse memes by applying commonsense reasoning and recalling mental examples of similar situations. This can be attributed to the *social commonsense* (Naslund et al., 2020; Arora et al., 2023; Office of the Surgeon General (OSG), 2023)[1] capabilities of humans which include *recognizing social norm violations* (e.g., hate speech, body shaming, misogyny, stereotyping, sexual content, vulgarity), *assessing credibility* (e.g., misinformation), *empathy and ethical judgment* (e.g., child exploitation, public decorum and privacy, cultural sensitivity, religious sensitivity), *contextual interpretation* (e.g., humor appropriateness), and *predicting consequences* (e.g., mental health impact, violence, substance abuse). This human-like capacity to interpret subtle or symbolic cues underscores the need for moderation frameworks that can replicate such higher-level reasoning rather than relying purely on text or raw pixels.

Early multimodal models have attempted to fuse vision and language through joint embeddings or cross-attention mechanisms (Shin and Narihira, 2021), yet they tend to place disproportionate emphasis on textual data. As a result, subtle image-based cues – such as historical references, cultural icons, or visually encoded irony – can slip through the cracks (Zhang et al., 2024). Even powerful contrastive models like CLIP (Radford et al., 2021) struggle when the meme's intent hinges on satire or understated visual hints requiring commonsense inference (Mazhar et al., 2025). These shortcomings highlight the urgent need for more holistic approaches that view images and text on equal footing, mirroring the way humans naturally interpret visual jokes and symbolic content.

A promising direction involves enriching model understanding through in-context examples (Liu et al., 2024) that illuminate both visual and textual nuances of a meme. Rather than processing an image in isolation, the model compares it against a small set of similar or thematically related images each annotated or tagged with the relevant commonsense insights needed for proper interpretation.

[1] https://en.wikipedia.org/wiki/Commonsense_reasoning

This strategy enables the model to draw parallels and detect patterns that might be missed if it were forced to rely on a single snapshot or textual prompt. By dynamically retrieving these curated examples alongside knowledge about harmful or deceptive imagery, the system gains the contextual backdrop necessary to catch everything from historical allusions to subtle visual sarcasm.

In this paper, we propose an adaptive in-context learning framework – MemeSense that synthesizes commonsense knowledge with semantically similar reference images to enhance the interpretation of meme content. Concretely, MemeSense retrieves a curated set of analogous memes, each annotated with cultural, historical, or situational context and incorporates these examples into a unified representation alongside the target meme. By embedding human-like commonsense cues directly into the model's input, we effectively steer its latent space toward the pertinent visual and textual signals present in the attached memes. This synergy allows the model to detect subtle or symbolic markers such as ironic juxtapositions, culturally coded imagery, or sarcastic overlays that often evade traditional pipelines. Our contributions are as follows.

- We develop a novel multi-staged framework to generate intervention for the harmful memes by leveraging cognitive shift vectors which reduce the requirement of demonstration examples during inference.
- We curate a wide-ranging dataset collection that emphasizes subtly harmful or text-scarce memes, filling a crucial gap in moderation research. This dataset lays the groundwork for a deeper exploration of nuanced meme analysis.
- Rigorous experiments demonstrate the efficacy of MemeSense even for the memes that do not contain any explicit text embedded in them as is usually the case. We obtain respectively 5% and 9% improvement in BERTScore over the most competitive baseline for the *memes with text* and the *memes without text*. Semantic similarity for memes with as well as without text (almost) doubles for MemeSense compared to the best baseline.

2 Related work

Visual in-context learning: In-context learning (ICL) has transformed LLM adaptation, enabling task generalization with few-shot demonstrations, and recent advancements have extended it to multimodal models for vision-language tasks like VQA (Brown et al., 2020; Alayrac et al., 2022). However, ICL suffers from computational inefficiency due to long input sequences and sensitivity to demonstration selection (Peng et al., 2024). To mitigate this, in-context vectors (ICVs) distill task-relevant information into compact representations, reducing the need for multiple examples (Hendel et al., 2023; Todd et al., 2024). While early non-learnable ICVs improved NLP efficiency, they struggled with multimodal tasks like VQA due to diverse vision-language inputs (Li et al., 2023; Yang et al., 2024). Recently, learnable ICVs dynamically capture essential task information, significantly enhancing VQA performance while lowering computational costs (Peng et al., 2024). These advancements underscore the importance of optimizing vector-based representations and refining ICL strategies to improve multimodal reasoning (Yin et al., 2024).

Intervention generation: Intervention strategies for online toxicity have largely focused on text-based issues like hate speech (Qian et al., 2019; Jha et al., 2024a), misinformation (He et al., 2023) and harm (Banerjee et al., 2024; Hazra et al., 2024; Banerjee et al., 2025), with limited exploration of multimodal challenges such as memes. While counterspeech interventions reshape discourse (Schieb and Preuss), their reliance on manual curation (Mathew et al., 2018) or supervised datasets limits scalability. Advances in LLMs and VLMs (Ghosh et al., 2024) have improved intervention capabilities but often lack contextual grounding, requiring knowledge-driven approaches (Dong et al., 2024). To address this, MemeGuard enhances meme interpretation using VLMs and knowledge ranking, enabling more precise and contextually relevant interventions (Jha et al., 2024a).

3 Methodology

In this work, we propose a framework that proceeds in three main stages – (a) **Stage I: Generation of commonsense parameters**: In Stage I, we generate commonsense parameters by instruction-tuning a multimodal large language model (MLLM) to predict contextually relevant insights for each image. (b) **Stage II: Selection of in-context exemplars**: We create a set of anchor images and retrieve corresponding in-context exemplars, which we later use in Stage III. (c) **Stage III: Learning cognitive shift**

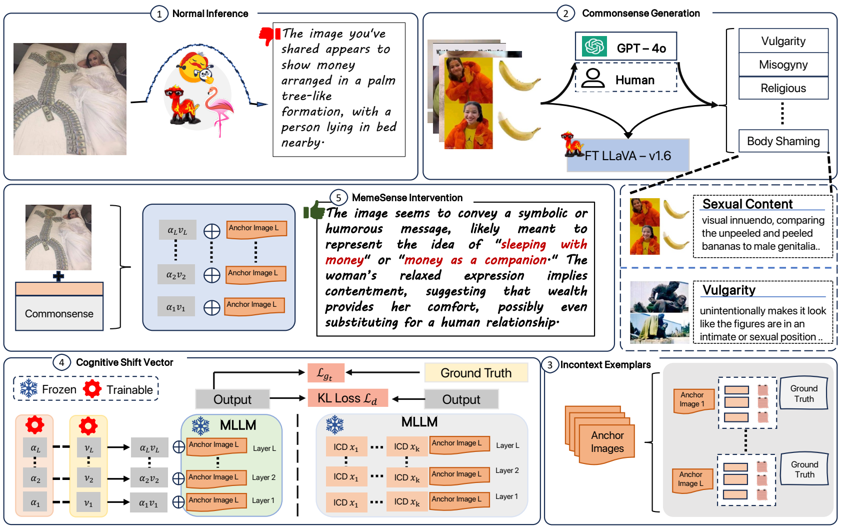

Figure 1: Schematic diagram of the MemeSense.

vector: Finally, we learn a cognitive shift vector by distilling general task information from the exemplars, and then guide the target model to align its representation with the insights derived from these exemplars. The overview of our proposed method is shown in Figure 1.

4 Preliminaries

A collection of images is denoted as $\mathcal{IMG}$, where each image img is an item of $\mathcal{IMG}$, i.e., $img \in \mathcal{IMG}$. GT_{img} describes the ground truth intervention on the image. In particular, GT_{img} contains the description about ***why the image can/can't be posted on social media?*** We consider a set of commonsense parameters $\mathscr{C}$ where i^{th} commonsense parameter is denoted as $c_i \in \mathscr{C}$. A pair consisting of an image and its corresponding commonsense parameters is denoted by $\langle img, \mathscr{C}_{img} \rangle$ where $\mathscr{C}_{img} \subseteq \mathscr{C}$. An image may be associated with multiple commonsense parameters. We partition $\mathcal{IMG}$ into two subsets: **(a)** the training set $\mathcal{IMG}_{tr}$, used at different stages of the training process, and **(b)** the test set $\mathcal{IMG}_{ts}$, reserved for evaluation. The set of training images $\mathcal{IMG}_{tr}$ and test images $\mathcal{IMG}_{ts}$ are disjoint, i.e., $\mathcal{IMG}_{tr} \cap \mathcal{IMG}_{ts} = \emptyset$. For **Stage I**, we build a training dataset $\mathcal{D}_{\mathscr{C}}$ consisting of images $\mathcal{IMG}_{tr}$ and their respective ground truth image description with commonsense parameters. We represent a fine-tuned vision language model with dataset $\mathcal{D}_{\mathscr{C}}$ as $\mathcal{M}_{\mathscr{C}}$. Further in **Stage II**, we construct an in-context (IC) learning set $\mathcal{D}_{\mathcal{IC}}$ (involves only images from $\mathcal{IMG}_{tr}$ set) to utilize in **Stage III** (see Section 4.3). Each instance in $\mathcal{D}_{\mathcal{IC}}$ is a tuple consisting of $\langle img_a, IC_{img}, GT_{img_a} \rangle$ where IC_{img} is the set of retrieved in-context examples of an anchor image img_a. Each in-context example consists of an image $img \neq img_a$, $\mathscr{C}_{img}$, GT_{img}. We define the cognitive shift vector set as $\mathcal{CSV}$ and the coefficient set as α. In **Stage III**, we use an instruction following MLLM as the target model ($\mathcal{M}$) to further generate the intervention defined as $\mathcal{M}_{ivt}$.

4.1 Stage I: Commonsense parameters

In this stage, we aim to fine-tune a vision-language model to produce relevant commonsense parameters for meme images. These parameters represent broad conceptual categories that help assess whether an image is *harmful*, *offensive*, or *inappropriate*, as discussed in (Arora et al., 2023; Office of the Surgeon General (OSG), 2023; Gongane et al., 2022). To create the training set $\mathcal{D}_{\mathscr{C}}$, we first use GPT-4o to automatically obtain commonsense parameters for $img \in \mathcal{IMG}_{tr}$ and then perform manual corrections. To evaluate GPT-4o's performance, we randomly select 40% of its responses and subject them to expert review by two professionals experienced in meme moderation. Their assessments revealed a strong consensus, with GPT-4o's outputs (96% agreement). Lastly, for the fine-tuning phase, we chose an open-weight model instead of a proprietary system such as GPT-4o, given the higher resource demands and stricter usage limitations. We employ the prompt in Table 10 to extract commonsense parameters from GPT-4o.
The example of commonsense parameters $\mathscr{C}_{img}$ of an image img is illustrated in Figure 2. By instruction tuning with this dataset $\mathcal{D}_{\mathscr{C}}$, we obtain the final vision-language model $\mathcal{M}_{\mathscr{C}}$, which we

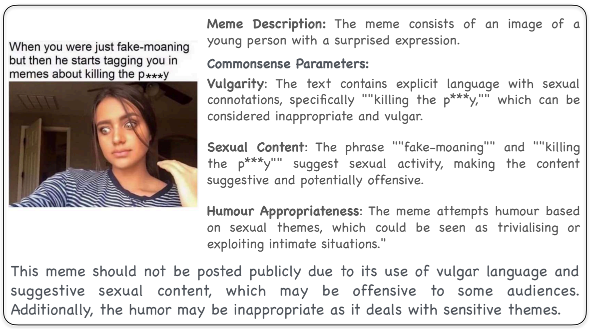

Figure 2: Representative example of a harmful meme and the annotated commonsense parameters along with intervention.

then use at inference time to produce commonsense parameters for new meme images.

Prompt for instruction tuning
[INST] <image>*Describe the meme and assess whether the meme has the potential to be perceived as vulgar, harmful, or unethical using specific commonsense parameters.* [\INST]

4.2 Stage II: Selection of in-context exemplars

In this stage, our objective is to create an in-context dataset $\mathcal{D}_{IC}$ that provides exemplars to guide the latent space of the target model in **Stage III**. To accomplish this, we reuse the training images $\mathcal{IMG}_{tr}$ and, following the authors in (Peng et al., 2024), treat each image $img \in \mathcal{IMG}_{tr}$ as an anchor. We denote an anchor image as img_a. We then select k in-context examples from $\mathcal{IMG}_{tr} \setminus img_a$ using multiple strategies. First, we randomly sample k candidate images to construct the set IC_{img} for each anchor. Apart from random selection, we also leverage semantic retrieval techniques that consider commonsense parameters, image representations, or a combination of both. The detailed setup of in-context retrieval is given in Section 6.

4.3 Stage III: Learning cognitive shift vectors

In this stage, the aim is to learn the trainable shift vector set $\mathcal{CSV}$ and coefficient set α so that the target model can generate proper intervention given an image img. We initialize a set of shift vectors $\mathcal{CSV} = \{csv^1, csv^2, \ldots, csv^L\}$ where each shift vector csv^ℓ corresponds to each layer $\ell \in L$ in the target model $\mathcal{M}$. L represents the number of layers in target model $\mathcal{M}$. Further, we consider a set of coefficients $\alpha = \{\alpha^1, \alpha^2, \ldots, \alpha^L\}$ which regulate the impact of these cognitive shift vectors across different layers in $\mathcal{M}$. After applying cognitive shift vector set $\mathcal{CSV}$ and α to the model $\mathcal{M}$, we obtain the final model as expressed in Equation 1.

$$\mathcal{M}^\ell_{ivt} = M^\ell + \alpha^\ell \cdot csv^\ell, \tag{1}$$

Following task analogies from (Peng et al., 2024), our objective is to align the output of $\mathcal{M}_{ivt}$ with the output obtained by including IC_{img} in model $\mathcal{M}$ for a given anchor image img_a. To achieve this, we minimize the KL divergence between the output distribution of $\mathcal{M}_{ivt}(img_a)$ and output distribution of $\mathcal{M}$ with IC exemplars IC_{img} for the anchor image img_a. The computation of $\mathscr{L}_{od}$ is given in Equation 2.

$$\begin{aligned}\mathscr{L}_{od} = KL\,(P(img_a|IC_{img};\mathcal{M}) \parallel \\ P(img_a|\mathcal{M}_{ivt}))\,.\end{aligned} \tag{2}$$

where $P(img_a|IC_{img};\mathcal{M})$ and $P(img_a|\mathcal{M}_{ivt})$ represent the output distribution of models $\mathcal{M}$ and $\mathcal{M}_{ivt}$ respectively for anchor image img_a.
Further we compute the intervention loss ($\mathscr{L}_{ivt}$) to make sure that the output of final model $\mathcal{M}_{ivt}(img_a)$ is aligned with the ground truth GT_{img_a} (see Equation 3)

$$\mathscr{L}_{ivt} = -\sum_{|\mathcal{D}_{IC}|} \log P(img_a|\mathcal{M}_{ivt}) \tag{3}$$

We compute the final loss as given in Equation 4. γ serves as a hyperparameter that determines the relative importance of output distribution loss and intervention loss.

$$\mathscr{L} = \mathscr{L}_{od} + \gamma \cdot \mathscr{L}_{ivt} \tag{4}$$

5 Datasets

To advance research on harmful meme intervention, we construct a novel dataset of implicitly harmful memes, sourced from various online social media platforms, including Facebook, Twitter, Instagram, and WhatsApp. Unlike existing datasets that primarily focus on memes with explicit textual content embedded in them, our dataset specifically targets memes that are implicitly harmful or lack embedded text (see Figure 3 for details). These cases pose additional challenges for AI models, as they require nuanced reasoning beyond surface-level textual analysis. Below, we detail our data collection and annotation process.
Data collection: We curate memes from publicly available online sources, including Facebook meme

(a) Harmful meme without text | (b) Harmful meme with text | (c) Non-harmful meme

Figure 3: Representative examples of different memes.

Commonsense category (meta)	Commonsense parameters	# Memes
Recognizing social norm violations	*Hate speech*	23
	Body shaming	74
	Misogyny	51
	Stereotyping	32
	Sexual content	105
	Vulgarity	135
Assessing credibility	*Misinformation*	4
Empathy and ethical judgements	*Child exploitation*	12
	Public decorum & Privacy	72
	Cultural sensitivity	60
	Religious sensitivity	14
Contextual interpretation	*Humor appropriateness*	251
Predicting consequences	*Mental health impact*	38
	Violence	43
	Substance abuse	7

Table 1: Distribution of various commonsense attributes.

pages[2], Twitter adult meme pages[3], public WhatsApp groups, and Instagram meme accounts[4]. In addition, we incorporate phallic[5]-themed memes[6] which may not appear overtly harmful at first glance but can carry implicit harmful implications when shared publicly. Our data collection process resulted in a total of 785 memes.

Filtering and annotation: To ensure relevance, we filter out memes that do not exhibit potential harm, specifically those that do not align with any of the 15 predefined commonsense harm categories (see Table 1). Two undergraduate annotators independently labeled each meme as ***harmful*** or ***non-harmful***. We retain only those memes that were unanimously marked as ***harmful*** by both annotators, resulting in a final dataset of 484 memes. Figure 3 illustrates representative examples of different memes from our collection. Once we have the annotations of the memes done we obtain the commonsense categories and the ground truth interventions for these memes using GPT-4o as was already discussed in Section 4.1.

[2]https://www.facebook.com/doublemean
[3]https://x.com/DefensePorn
[4]https://www.instagram.com/stoned_age_humour
[5]https://en.wikipedia.org/wiki/Phallus
[6]https://humornama.com/memes/penis-memes/

5.1 ICMM data

In addition to our curated dataset, we also consider the publicly available *Intervening Cyberbullying in Multimodal Memes* (ICMM) dataset (Jha et al., 2024a) for evaluation of our approach. This dataset consists of 1000 cyberbullying memes along with their corresponding crowdsourced interventions. After filtering out the corrupted images, we obtain a set of 985 memes along with their ground truth interventions.

6 Experimental setup

This section discusses the different experimental configurations of MemeSense.

6.1 Baselines

We evaluate our proposed approach against several baseline methods, including state-of-the-art meme intervention techniques and various prompting strategies.

(1) MemeGuard (Jha et al., 2024a): We adopt MemeGuard, a state-of-the-art meme intervention generation model, as a baseline. Given a meme, we use a VLM to generate five descriptive answers. To filter out irrelevant content, we compute the semantic similarity between the input meme and the generated sentences, retaining only those exceeding a 0.2 threshold (determined via manual inspection). Finally, another VLM generates the intervention based on the meme and the filtered descriptions.

(2) MemeMQA (Modified) (Agarwal et al., 2024): We extend the MemeMQA framework for intervention generation by removing its target identification module and repurposing its explanation generation module. Originally designed to identify targets in hateful memes and explain predictions, MemeMQA now directly generates interventions.

(3) Commonsense-enhanced prompting: Given a meme and its automatically generated commonsense parameters, the VLM is instructed to generate an intervention.

(4) In-context learning (ICL) (Zeng et al., 2024): For a given target meme, we select k ($\in \{1, 2, 4\}$) demonstration examples from the training set, including their annotated commonsense, and provide them as context before prompting the VLM to generate an intervention. For the selection of in-context examples, we use random and semantic retrieval techniques similar to **Stage II** (Section 4.2).

6.2 MemeSense framework

Recall that MemeSense consists of three major stages leveraging (I) multimodal LLMs for generation of commonsense parameter, (II) in-context exemplars selection and (III) subsequent learning of the cognitive shift vector for the **intervention generation**.

For the **Stage I**, we utilize the `llava-v1.6-mistral-7b-hf`[7] model, fine-tuned with QLoRA (Dettmers et al., 2023) over 10 epochs using a batch size of 16 and a learning rate of 2×10^{-4}, with weight decay for optimization.

For the **Stage II**, We employ various strategies for selecting in-context exemplars, detailed as follows:

Commonsense-based retrieval: For each predefined commonsense parameter, we select up to five instances from our training set to form a lookup set. Given an anchor image img and its corresponding annotated commonsense parameters, we iteratively retrieve at least one instance per parameter to construct the k demonstration examples.

Image-based retrieval: For a given anchor image img, we retrieve k demonstrations by computing their semantic similarity with img from the training subset. To achieve this, we first encode all images into dense vector representations using the `CLIP-ViT`[8] multimodal embedding model. When an anchor image is provided as a query, we map it into the same vector space, enabling an efficient similarity search. We then perform Approximate Nearest Neighbor (ANN) (Wang et al., 2021) search to identify the top k most similar images. Their corresponding commonsense parameters and ground truth interventions are retrieved as in-context examples, ensuring a contextually relevant selection.

Combined retrieval: We also experiment with constructing the k in-context demonstrations by combining the above two approaches. Here, we select c instances from the commonsense based retrieval and $(k-c)$ instances from the image-based retrieval, where $c \in \{1, 2\}$.

For **Stage III**, we primarily employ the `idefics2-8B-base`[9] model to learn cognitive shift vectors and perform inference. In addition, we explore `idefics-9B`[10] and `OpenFlamingo`[11] for intervention generation (results presented in Appendix E). The number of in-context demonstration examples is one of {1, 2, 4}, maintaining a fixed batch size of 2. The shift vector undergoes training for 10 epochs to ensure effective adaptation and we choose γ as 0.5.

6.3 Baseline models

For baselines involving zero-shot prompting and in-context learning (ICL), we leverage the same aligned MLLMs used in MemeSense – `idefics2-8B`, `idefics-9B`, and `OpenFlamingo` – for intervention generation.

The **MemeQA** baseline adopts a dual-model architecture, comprising:

(1) An MLLM for rationale generation, aligned with the MemeSense models.

(2) A `T5-large` model for intervention generation.

The rationale generation MLLM is fine-tuned for one epoch with a batch size of 4 and a learning rate of 5×10^{-5}.

MemeGuard, another baseline, employs two MLLMs for intervention generation, using models aligned with those in MemeSense to ensure consistency in evaluation.

6.4 Evaluation metrics

To rigorously assess the quality of generated interventions, we employ a diverse set of evaluation metrics spanning semantic similarity, lexical accuracy, and readability. Semantic metrics such as BERTScore (Zhang* et al., 2020) and semantic cosine similarity (Rahutomo et al., 2012) measure the alignment between generated and reference interventions in embedding space. Lexical metrics, including ROUGE-L (Lin, 2004) and BLEU-4 (Papineni et al., 2002), evaluate surface-level text overlap and n-gram precision. Further, a readability score assesses fluency and ease of comprehension, ensuring the interventions are not only accurate but

[7] `https://huggingface.co/llava-hf/llava-v1.6-mistral-7b-hf`

[8] `sentence-transformers/clip-ViT-B-32`

[9] `https://huggingface.co/HuggingFaceM4/idefics2-8b-base`

[10] `https://huggingface.co/HuggingFaceM4/idefics-9b`

[11] `https://huggingface.co/openflamingo/OpenFlamingo-9B-vitl-mpt7b`

also coherent and accessible. This holistic evaluation framework enables a nuanced assessment of intervention effectiveness across multiple linguistic dimensions.

7 Results

We structure our experimental results into three key sections. First, we present insights derived from our dataset, highlighting key patterns and observations. Next, we evaluate the performance of our framework on the ICMM dataset, examining its effectiveness in generating interventions. Finally, we delve into a detailed breakdown of performance across different commonsense meta-categories, offering a deeper understanding of the model's strengths and limitations in various contexts.

Result for our dataset: In Tables 2 and 3, we compare the performance of our framework, MemeSense, with various baselines on memes without text and memes with text, respectively. Across both settings, MemeSense (combined) consistently achieves the highest values for BERTScore (0.91), semantic similarity (0.71 for the memes without text, 0.78 for text-based memes), and ROUGE-L (0.35 and 0.37, respectively), demonstrating its superior capability in generating semantically meaningful and contextually appropriate responses. Among the baseline methods, commonsense-anchored ICL performs competitively but lags behind MemeSense, particularly in terms of semantic similarity score, highlighting the importance of hybrid reasoning strategies.

For memes without text, direct prompting methods struggle with low semantic similarity (≤ 0.3), while MemeSense (combined) significantly outperforms them (semantic similarity = 0.71). Similarly, for memes with text, MemeSense achieves notable improvements in both semantic alignment and lexical overlap (BLEU: 0.08–0.09), reflecting its ability to effectively integrate commonsense and image-grounded reasoning. Overall, these results demonstrate that the MemeSense (combined) approach integrating image-anchored, and commonsense-anchored in-context learning (ICL), effectively enhances reasoning and interpretation across different meme types.

Result for ICMM data: In Table 4, we show the result of various baselines and compare them with MemeSense for the ICMM dataset. Direct prompting achieves the highest readability (67.02) but performs poorly in semantic alignment (SeSS = 0.15, ROUGE-L = 0.03, BLEU = 0.001), while adding commonsense knowledge reduces readability further (52.34) without improving semantic scores. In-context learning (ICL) methods, including random, image-anchored, and commonsense-anchored ICL, improve semantic similarity (0.16–0.22) and ROUGE-L (0.09–0.1) but suffer from significantly lower readability (19.63–25.38). Among meme-specific baseline models, **MemeQA** performs best (SeSS = 0.24, readability = 54.45) as it requires explicit training, while **MemeGuard** underperforms across all metrics (SeSS = 0.18, readability = 34.45). MemeSense outperforms all baselines, with MemeSense (commonsense anchored ICL) achieving strong semantic alignment (SeSS = 0.27), while MemeSense (combined) emerges as the best overall method with the highest BERTScore (0.875) and SeSS (0.31), reasonable readability (45.57), and competitive ROUGE-L (0.11) and BLEU (0.008) scores. This suggests that structured multimodal approaches, particularly MemeSense (combined), provide the best balance between semantic coherence and fluency, making it the most effective meme intervention generation strategy.

Results for social commonsense categories: Table 5 presents the performance of our model across different broad social commonsense categories, evaluated using BERTScore (F1), semantic similarity (SeSS), and ROUGE-L. Notably, for all four categories, the results are very similar showing the robustness of the design of MemeSense. The model achieves the highest scores in *recognizing social*

Method	*BERTScore (F1)*	*SeSS*	*Readability*	*ROUGE-L (Avg)*	*BLEU (Avg)*
Direct prompting	0.81	0.27	**53.36**	0.05	0.001
Direct prompting *(w. commonsense)*	0.81	0.3	21.55	0.05	0.002
Random ICL	0.87	0.49	35.06	0.19	0.01
Image anchored ICL	0.86	0.41	36.49	0.17	0.02
Commonsense anchored ICL	0.88	0.46	34.12	0.18	0.02
MemeMQA	0.86	0.51	52.86	0.08	0.008
MemeGuard	0.82	0.35	51.69	0.09	0.005
MemeSense *(random ICL)*	0.9	0.68	46.22	0.34	0.07
MemeSense *(image anchored ICL)*	0.9	0.7	45.57	0.35	0.08
MemeSense *(commonsense anchored ICL)*	0.91	0.7	45.65	0.35	**0.09**
MemeSense *(combined)*	**0.91**	**0.71**	44.07	**0.35**	0.08

Table 2: Result for memes without text. ***SeSS***: semantic similarity.

Method	*BERTScore (F1)*	*SeSS*	*Readability*	*ROUGE-L (Avg)*	*BLEU (Avg)*
Direct prompting	0.81	0.35	**54.59**	0.04	0.001
Direct prompting *(w. commonsense)*	0.8	0.28	22.02	0.04	0.001
Random ICL	0.86	0.52	31.94	0.18	0.02
Image anchored ICL	0.87	0.49	31.52	0.18	0.02
Commonsense anchored ICL	0.88	0.55	33.25	0.19	0.03
MemeQA	0.86	0.54	50.28	0.1	0.009
MemeGuard	0.84	0.39	36.36	0.09	0.004
MemeSense *(random ICL)*	0.91	0.77	46.64	0.36	0.08
MemeSense *(image anchored ICL)*	0.91	0.77	44.33	0.35	0.07
MemeSense *(commonsense anchored ICL)*	0.91	0.78	48.74	**0.38**	**0.09**
MemeSense *(combined)*	**0.91**	**0.78**	43.38	0.37	0.08

Table 3: Result for memes with text. ***SeSS***: semantic similarity.

Method	*BERTScore (F1)*	*SeSS*	*Readability*	*ROUGE-L (Avg)*	*BLEU (Avg)*
Direct prompting	0.8	0.15	67.02	0.03	0.001
Direct prompting with commonsense	0.8	0.14	52.34	0.03	0.004
Random ICL	0.82	0.16	19.63	0.09	0.005
Image anchored ICL	0.82	0.2	22.16	0.1	0.006
Commonsense anchored ICL	0.84	0.22	25.38	0.1	0.006
MemeQA	0.85	0.24	54.45	0.1	0.007
MemeGuard	0.79	0.18	34.45	0.04	0.001
MemeSense *(random ICL)*	0.84	0.18	44.03	0.11	0.007
MemeSense *(image anchored ICL)*	0.85	0.25	42.79	0.1	0.007
MemeSense *(commonsense anchored ICL)*	0.86	0.27	42.22	0.11	0.009
MemeSense *(combined)*	**0.87**	**0.31**	45.57	0.11	0.008

Table 4: Result for the ICMM dataset.

Meta category (Commonsense)	*BERTScore (F1)*	*SeSS*	*ROUGE-L (Avg)*
Contextual interpretation	0.91	0.78	0.37
Empathy and ethical judgements	0.90	0.75	0.33
Predicting consequences	0.90	0.72	0.33
Recognizing social norm violations	**0.91**	**0.79**	**0.38**

Table 5: Meta category-wise evaluation results.

norm violations (BERTScore: 0.91, SeSS: 0.79, ROUGE-L: 0.38), suggesting strong alignment with human references in identifying and intervening in socially inappropriate memes containing themes such as *vulgarity*, *sexual content* etc. For the other three categories also the results are quite close in terms of all three metrics (BERTScore: 0.90/0.91, SeSS: 0.72-0.78, ROUGE-L: 0.33-0.37).

8 Discussion

Error analysis: To better understand the limitations of MemeSense, we conduct a detailed error analysis by examining its predictions and identifying cases where erroneous classifications occur. We categorize the errors into two types:
(1) *False negative* (Category 1 error): Instances where the meme is actually harmful and should be flagged as unsafe, but MemeSense incorrectly predicts it as safe for posting.
(2) *Improper reasoning* (Category 2 error): Cases where the model correctly identifies the meme as unsafe but provides incorrect or inadequate reasoning for its decision.
Our analysis focuses on memes without explicit text, where reasoning relies primarily on visual cues. Among 51 such instances in our dataset, MemeSense exhibits Category 1 errors in 6 cases. Notably, in 5 out of these 6 cases, the commonsense parameter generation stage fails to accurately infer the harmful category, leading to incorrect classification. A specific example of this failure is observed when the model incorrectly identifies *cultural sensitivity* as the primary harmful category for a meme that is actually *vulgar*, ultimately leading to its misclassification as safe for posting.
Further, we identify one instance of Category 2 error, where the model predicts the meme as unsafe but fails to provide a coherent justification. This error arises due to improper reasoning during the commonsense parameter generation stage, which affects the interpretability and reliability of the model's intervention.
Ablation studies: In the error analysis, we observed the major prediction error appeared due to the incorrect generation of commonsense parameters. Hence we investigate, how much the final inference is dependent on the generated commonsense parameters. To achieve this, we obtain the inference from our approach without providing commonsense information to the model. Using only the input image and its corresponding description, we attempt to infer the intervention from our approach using the best method (MemeSense (combined)). The combined model is trained using the commonsense information. However, during the inference we are not providing the commonsense, removing the requirement of commonsense generation module during inference. We observe a maximum decline in semantic similarity score of 4% without commonsense information. In addition, we observe that the interventions are more descriptive, which is reflected in the increase of the *readability* score.

Test set	*BERTScore (F1)*	*SeSS*	*Readability*	*ROUGE-L (Avg)*	*BLEU (Avg)*
Memes without text	0.89(**-.02**)	0.68(**-.03**)	51.02(**+6.95**)	0.31(**-.04**)	0.07(**-.01**)
Memes with text	0.9 (**-.01**)	0.74(**-.04**)	47.79(**+4.41**)	0.32(**-.04**)	0.06(**-.02**)
ICMM	0.85(**-.02**)	0.27(**-.04**)	54.19(**+8.62**)	0.10(**-.01**)	0.007(**-.001**)

Table 6: Result for intervention generation for different test sets without using the commonsense parameters.

9 Conclusion

In this work, we introduced MemeSense, a three-stage, adaptive in-context learning framework that integrates visual and textual cues with social commonsense knowledge for robust meme moderation. By combining compact latent representations, carefully retrieved in-context exemplars, and cognitive shift vectors, our approach captures subtle, implicitly harmful signals, including memes without explicit text that often evade traditional pipelines. Experiments on our curated dataset and the *ICMM* benchmark highlight MemeSense's superior performance in generating semantically aligned interventions, surpassing state-of-the-art baselines. We hope MemeSense inspires broader research in in-context learning toward fostering safer, more responsible online communities.

10 Limitation

A principal limitation of MemeSense lies in its reliance on automatically generated commonsense parameters and curated in-context exemplars, which may not capture the full spectrum of cultural or linguistic nuances. In particular, memes containing highly context-dependent references or adversarial manipulations could circumvent the system's current retrieval and commonsense inference components. In addition, because the approach depends on a finite set of annotated harmful categories, novel or emerging social norms might remain undetected. Addressing these concerns through broader annotation schemas, continuous model adaptation, and more nuanced retrieval strategies constitutes a pivotal direction for future work.

11 Ethical considerations

This work aims to promote safer online environments by detecting and mitigating harmful or offensive memes. However, automated moderation tools, including MemeSense, carry risks of over-moderation, potentially limiting free expression, especially in contexts where satire or cultural references are misinterpreted. We strived to minimize biases in data collection and annotation by involving diverse annotators and ensuring consistent labeling protocols. Yet, subjective judgments on harmfulness may still reflect annotators' cultural and personal perspectives. Moreover, the system's performance hinges on training data quality, introducing the possibility of inadvertently perpetuating harmful societal biases. Transparent reporting of system limitations and the use of MemeSense as a supplementary tool rather than a definitive arbiter remain crucial in safeguarding fairness and accountability in online content moderation.

References

Siddhant Agarwal, Shivam Sharma, Preslav Nakov, and Tanmoy Chakraborty. 2024. Mememqa: Multimodal question answering for memes via rationale-based inferencing. *Preprint*, arXiv:2405.11215.

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katie Millican, 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. *Preprint*, arXiv:2204.14198.

Rie Kubota Ando and Tong Zhang. 2005. A framework for learning predictive structures from multiple tasks and unlabeled data. *Journal of Machine Learning Research*, 6:1817–1853.

Arnav Arora, Preslav Nakov, Momchil Hardalov, Sheikh Muhammad Sarwar, Vibha Nayak, Yoan Dinkov, Dimitrina Zlatkova, Kyle Dent, Ameya Bhatawdekar, Guillaume Bouchard, and Isabelle Augenstein. 2023. Detecting harmful content on online platforms: What platforms need vs. where research efforts go. *ACM Comput. Surv.*, 56(3).

Somnath Banerjee, Sayan Layek, Hari Shrawgi, Rajarshi Mandal, Avik Halder, Shanu Kumar, Sagnik Basu, Parag Agrawal, Rima Hazra, and Animesh Mukherjee. 2025. Navigating the cultural kaleidoscope: A hitchhiker's guide to sensitivity in large language models. *Preprint*, arXiv:2410.12880.

Somnath Banerjee, Sayan Layek, Soham Tripathy, Shanu Kumar, Animesh Mukherjee, and Rima Hazra. 2024. Safeinfer: Context adaptive decoding time safety alignment for large language models. *Preprint*, arXiv:2406.12274.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. *Preprint*, arXiv:2005.14165.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. Qlora: Efficient finetuning of quantized llms. *Preprint*, arXiv:2305.14314.

Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Jingyuan Ma, Rui Li, Heming Xia, Jingjing Xu, Zhiyong Wu, Tianyu Liu, Baobao Chang, Xu Sun, Lei Li, and Zhifang Sui. 2024. A survey on in-context learning. *Preprint*, arXiv:2301.00234.

Akash Ghosh, Arkadeep Acharya, Sriparna Saha, Vinija Jain, and Aman Chadha. 2024. Exploring the frontier of vision-language models: A survey of current methodologies and future directions. *Preprint*, arXiv:2404.07214.

Vaishali U Gongane, Mousami V Munot, and Alwin D Anuse. 2022. Detection and moderation of detrimental content on social media platforms: current status and future directions. *Social Network Analysis and Mining*, 12(1):129.

Rima Hazra, Sayan Layek, Somnath Banerjee, and Soujanya Poria. 2024. Safety arithmetic: A framework for test-time safety alignment of language models by steering parameters and activations. *Preprint*, arXiv:2406.11801.

Bing He, Mustaque Ahamad, and Srijan Kumar. 2023. Reinforcement learning-based counter-misinformation response generation: A case study of covid-19 vaccine misinformation. In *Proceedings of the ACM Web Conference 2023*, WWW '23, page 2698–2709, New York, NY, USA. Association for Computing Machinery.

Roee Hendel, Mor Geva, and Amir Globerson. 2023. In-context learning creates task vectors. *Preprint*, arXiv:2310.15916.

Raghav Jain, Krishanu Maity, Prince Jha, and Sriparna Saha. 2023. Generative models vs discriminative models: Which performs better in detecting cyberbullying in memes? In *2023 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8.

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

Prince Jha, Krishanu Maity, Raghav Jain, Apoorv Verma, Sriparna Saha, and Pushpak Bhattacharyya. 2024b. Meme-ingful analysis: Enhanced understanding of cyberbullying in memes through multimodal explanations. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 930–943, St. Julian's, Malta. Association for Computational Linguistics.

Li Li, Jiawei Peng, Huiyi Chen, Chongyang Gao, and Xu Yang. 2023. How to configure good in-context sequence for visual question answering. *Preprint*, arXiv:2312.01571.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Sheng Liu, Haotian Ye, Lei Xing, and James Zou. 2024. In-context vectors: Making in context learning more effective and controllable through latent space steering. *Preprint*, arXiv:2311.06668.

Krishanu Maity, A. S. Poornash, Shaubhik Bhattacharya, Salisa Phosit, Sawarod Kongsamlit, Sriparna Saha, and Kitsuchart Pasupa. 2024. Hatethaisent: Sentiment-aided hate speech detection in thai language. *IEEE Transactions on Computational Social Systems*, 11(5):5714–5727.

Binny Mathew, Navish Kumar, Ravina, Pawan Goyal, and Animesh Mukherjee. 2018. Analyzing the hate and counter speech accounts on twitter. *Preprint*, arXiv:1812.02712.

Abdullah Mazhar, Zuhair hasan shaik, Aseem Srivastava, Polly Ruhnke, Lavanya Vaddavalli, Sri Keshav Katragadda, Shweta Yadav, and Md Shad Akhtar. 2025. Figurative-cum-commonsense knowledge infusion for multimodal mental health meme classification. *Preprint*, arXiv:2501.15321.

John A Naslund, Ameya Bondre, John Torous, and Kelly A Aschbrenner. 2020. Social media and mental health: Benefits, risks, and opportunities for research and practice. *Journal of Technology in Behavioral Science*, 5(3):245–257.

Office of the Surgeon General (OSG). 2023. *Social Media and Youth Mental Health: The U.S. Surgeon General's Advisory*. US Department of Health and Human Services, Washington, DC.

OpenAI, :, Aaron Hurst, Adam Lerer, and Adam P. Goucher et al. 2024. Gpt-4o system card. *Preprint*, arXiv:2410.21276.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Yingzhe Peng, Chenduo Hao, Xu Yang, Jiawei Peng, Xinting Hu, and Xin Geng. 2024. Live: Learnable in-context vector for visual question answering. *Preprint*, arXiv:2406.13185.

Jing Qian, Anna Bethke, Yinyin Liu, Elizabeth Belding, and William Yang Wang. 2019. A benchmark dataset for learning to intervene in online hate speech. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4755–4764, Hong Kong, China. Association for Computational Linguistics.

Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. 2025. Qwen2.5 technical report. *Preprint*, arXiv:2412.15115.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. *Preprint*, arXiv:2103.00020.

Faisal Rahutomo, Teruaki Kitasuka, and Masayoshi Aritsugi. 2012. Semantic cosine similarity.

Punyajoy Saha, Aalok Agrawal, Abhik Jana, Chris Biemann, and Animesh Mukherjee. 2024. On zero-shot counterspeech generation by LLMs. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 12443–12454, Torino, Italia. ELRA and ICCL.

Carla Schieb and Mike Preuss. Governing hate speech by means of counterspeech on facebook.

Shivam Sharma, Ramaneswaran S, Udit Arora, Md. Shad Akhtar, and Tanmoy Chakraborty. 2023. MEMEX: Detecting explanatory evidence for memes via knowledge-enriched contextualization. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5272–5290, Toronto, Canada. Association for Computational Linguistics.

Andrew Shin and Takuya Narihira. 2021. Transformer-exclusive cross-modal representation for vision and language. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 2719–2725, Online. Association for Computational Linguistics.

Christian Stab, Tristan Miller, Benjamin Schiller, Pranav Rai, and Iryna Gurevych. 2018. Cross-topic argument mining from heterogeneous sources. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3664–3674, Brussels, Belgium. Association for Computational Linguistics.

Gemini Team, Rohan Anil, and Sebastian Borgeaud et al. 2024. Gemini: A family of highly capable multimodal models. *Preprint*, arXiv:2312.11805.

Eric Todd, Millicent L. Li, Arnab Sen Sharma, Aaron Mueller, Byron C. Wallace, and David Bau. 2024. Function vectors in large language models. *Preprint*, arXiv:2310.15213.

Mengzhao Wang, Xiaoliang Xu, Qiang Yue, and Yuxiang Wang. 2021. A comprehensive survey and experimental comparison of graph-based approximate nearest neighbor search. *Preprint*, arXiv:2101.12631.

Xu Yang, Yongliang Wu, Mingzhuo Yang, Haokun Chen, and Xin Geng. 2024. Exploring diverse in-context configurations for image captioning. *Preprint*, arXiv:2305.14800.

Shukang Yin, Chaoyou Fu, Sirui Zhao, Ke Li, Xing Sun, Tong Xu, and Enhong Chen. 2024. A survey on multimodal large language models. *National Science Review*, 11(12).

Yuchen Zeng, Wonjun Kang, Yicong Chen, Hyung Il Koo, and Kangwook Lee. 2024. Can mllms perform text-to-image in-context learning? *Preprint*, arXiv:2402.01293.

Jingyi Zhang, Jiaxing Huang, Sheng Jin, and Shijian Lu. 2024. Vision-language models for vision tasks: A survey. *Preprint*, arXiv:2304.00685.

Tianyi Zhang*, Varsha Kishore*, Felix Wu*, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

A Prompts

The prompt for generating ground truth commonsense parameters and intervention using GPT-4o is represented in the Table 10.

Test set	*BERTScore (F1)*	*SeSS*	*Readability*	*ROUGE-L (Avg)*	*BLEU (Avg)*
Memes without text	0.88(**-.03**)	0.64(**-.07**)	36.76(**-7.31**)	0.27(**-.08**)	0.05(**-.03**)
Memes with text	0.89(**-.02**)	0.69(**-.09**)	46.36(**+2.98**)	0.28(**-.08**)	0.05(**-.03**)
ICMM	0.85(**-.02**)	0.27(**-.04**)	34.07(**-11.50**)	0.10(**-.01**)	0.007(**-.001**)

Table 7: Result for intervention generation for different test sets using randomly selected commonsense parameters.

Used model	**Method**	**BERTScore (F1)**	**SeSS**	**Rouge-L (Avg)**
		Memes without text		
`Idefics-9B`	**MemeSense** (*random ICL*)	0.89	0.69	0.31
	MemeSense (*combined ICL*)	0.9	0.71	0.34
`OpenFlamingo-9B`	**MemeSense** (*random ICL*)	0.88	0.67	0.29
	MemeSense (*combined ICL*)	0.9	0.7	0.32
		Memes with text		
`Idefics-9B`	**MemeSense** (*random ICL*)	0.9	0.75	0.33
	MemeSense (*combined ICL*)	0.91	0.77	0.36
`OpenFlamingo-9B`	**MemeSense** (*random ICL*)	0.89	0.74	0.32
	MemeSense (*combined ICL*)	0.91	0.77	0.35
		ICMM data		
`Idefics-9B`	**MemeSense** (*random ICL*)	0.85	0.27	0.1
	MemeSense (*combined ICL*)	0.86	0.3	0.1
`OpenFlamingo-9B`	**MemeSense** (*random ICL*)	0.85	0.26	0.09
	MemeSense (*combined ICL*)	0.85	0.29	0.1

Table 8: Comparative results of MemeSense using other models.

B MemeSense sensitivity analysis

In addition to the ablation studies presented in Table 6, we conduct a sensitivity analysis to assess the impact of variations in the commonsense information provided to the model. Specifically, we evaluate how MemeSense (combined) performs when supplied with randomly selected commonsense knowledge during inference. This experiment aims to understand the model's sensitivity to incorrect or unrelated commonsense attributes.
As shown in Table 7, we observe a noticeable decline in performance across key metrics when randomly selected commonsense information is used. In particular, the semantic similarity score decreases by approximately 9%, indicating that misattributed commonsense knowledge can significantly affect the model's final outcome. The decline is also reflected in BERTScore, ROUGE-L, and BLEU, demonstrating the reliance of MemeSense on relevant commonsense reasoning for effective

intervention generation. Interestingly, readability exhibits a slight improvement for memes with text, which could be attributed to the increased linguistic diversity introduced by the random commonsense selection. These findings highlight the importance of precise commonsense attribution in ensuring robust and reliable meme interpretation.

C Intervention quality measurement

Measuring argument quality. We aim to measure the argument characteristic of the generated response commonly used for measuring quality of online *counterspeech* (Saha et al., 2024). We use a `roberta-base-uncased` model[12] finetuned on the argument dataset (Stab et al., 2018). Given this model, we pass each generated intervention through the classifier to predict a confidence score, which would denote the argument quality. We obtain confidence scores of 0.67, 0.74, 0.79 for the memes without texts, memes with text, and the ICMM dataset respectively suggesting strong argument quality of the generated interventions.

Correlation with human judgments. While we present most of our results with automatic metrics, it is important to understand if they correlate with human judgments. We took two metrics – BERTScore (F1) and ROUGE-L (Avg). For each metric, we randomly extract 25 samples from the prediction set. We present these to human annotators (researchers in this domain) and ask them to rate the quality of intervention from 1-5, 5 being the best and 1 being the worst. The Spearman's rank correlations between the human judgments (ordinal) and the automated metrics (continuous) are 0.58 and 0.49 respectively which indicates moderate to high correlation[13]. Given the subjective nature of the task, these results highlight a substantial consistency between automated metrics and human judgments, affirming their reliability.

D Additional experimental settings

D.1 Baselines

In Table 12 we demonstrate the different prompts used for the **MemeGuard** baseline. The hyperparameters for the experiments with this baseline are noted in Table 13.

[12] `https://huggingface.co/chkla/roberta-argument`

[13] `https://datatab.net/tutorial/spearman-correlation`

E Results for different models

In the Table 8, we show the comparative results of MemeSense using different models (`Idfics-9B` and `OpenFlamingo-9B`). Here we use the annotated data mentioned in 4.1, and the retrieval of in-context exemplars mentioned in the Section 4.2 to train the cognitive shift vectors (mentioned in the Section 4.3) using the two models. Then we perform the inference using trained cognitive shift vectors. We observe a similar pattern with these two models. Morever, `Idefics-9B` showing overall superior performance than `OpenFlamingo-9B`.

			
Direct Prompting	The meme is a popular form of communication that has been around for centuries. It is a way to express ideas, opinions, and emotions through images and text. Memes can be funny, serious, or even offensive. They can be used to spread information, make a point, or just entertain.	The meme is a humorous image that depicts a woman pouring milk into a container. The text on the image reads, “Her: “Just make sure you don't finish inside of me.” Me inside of her 2 minutes ago.” The meme is a play on the phrase “just make sure you don't finish inside of me,” which is often used to indicate that a person should use a condom during sex. The meme is humorous because it suggests that the woman is concerned about the man finishing inside of her, but the man is actually inside of her at the time.	The meme is a picture of a man with the caption “when the school shooter finally leaves and the hottest girl in class is laying dead on the floor”. The meme is making fun of the school shooter and the girl who is dead. The meme is also making fun of the fact that the girl is dead.
Direct Prompting (w. commonsense)	In the image, a person is sitting on a chair, holding a hot dog on a stick The hot dog is being cooked on a grill, and the person is enjoying the process There are two other hot dogs on the grill, one located closer to the person and the other further away In addition to the hot dogs, there are two marshmallows on the grill, one near the person and the other further away The overall atmosphere appears to be a fun and relaxed outdoor gathering.	The meme is inappropriate and violates community guidelines against explicit or adult content. It is not suitable for general audiences and may be offensive or uncomfortable to some viewers.	The meme is a screenshot of a man smiling and posing for the camera. The text on the image reads "When the school shooter finally leaves and the hottest girl in class is laying dead on the floor its free real estate". The combination of the text and image attempts to make a joke about a potentially tragic situation by using dark humor. The meme references a school shooting a highly sensitive and traumatic subject. Making jokes about such an event can be perceived as trivializing a grave and distressing experience potentially causing harm to those affected.
Random ICL	Celebrating individuality and embracing diversity enhances understanding and acceptance among us all.	This meme is vulgar, inappropriate, and potentially harmful, as it promotes a misogynistic view of sexual consent.	The meme should not be posted publicly due to its insensitive and inappropriate content, which could offend or disturb certain audiences.
Image anchored ICL	This meme humorously highlights the playful creativity and shared sense of humor in relationships, encouraging laughter and bonding over shared jokes.	This meme is vulgar, inappropriate, and potentially harmful, as it promotes a misogynistic view of sexual consent.	The meme should not be posted publicly as it trivializes the trauma of school shootings and can be distressing to individuals who have experienced similar events. It is inappropriate to use humor in this context, as it can be perceived as insensitive and offensive.
Commonsense anchored ICL	The image features a man holding a hot dog on a stick, with a marshmallow on one end and a penis on the other This meme is disrespectful and inappropriate, as it makes light of a serious issue, such as erectile dysfunction, and uses it to create a humorous or entertaining situation. It is important to be aware of the impact of such memes and to promote respectful and inclusive communication.	The meme should not be posted publicly due to its vulgar and suggestive content, which could be offensive to many viewers.	The meme appears to be inappropriate for public posting due to its insensitive and offensive content. It trivializes the trauma of school shootings and could be distressing to individuals who have experienced similar events. The use of humor in this context may be perceived as insensitive and offensive, particularly considering the serious nature of school shootings. The subject matter could be perceived as a violation of public sensitivity and a discussion of serious topics in a public context.
MemeQA	The meme should be restricted due to its content and content. It is inappropriate for public posting and should not be shared publicly.	The meme should be restricted due to its content and inappropriate content.	The meme should not be posted publicly due to its potential to be offensive to some audiences.
MemeGuard	The meme is making a claim that the person is having a barbecue with a hot dog and marshmallows, but instead of a traditional hot dog, there is a penis-shaped hot dog. The meme is likely meant to be humorous and playful, as it takes a common outdoor activity like grilling and adds a provocative twist to it.	The meme is harmful because it reinforces negative stereotypes about women and their sexuality. It suggests that women are concerned about pregnancy and STIs, and that they are not interested in sexual pleasure. It also suggests that men are not concerned about these issues and are only interested in their own pleasure. This is harmful because it perpetuates the idea that women are not equal to men in terms of sexuality and that they are not entitled to sexual pleasure.	This meme contains highly inappropriate and offensive content, trivializing school shootings and making an unethical implication about victims. This kind of meme should be critically examined and condemned rather than shared
MemeSense (Random ICL)	The meme should be restricted from public posting due to its vulgar and suggestive nature, which could be perceived as offensive or inappropriate.	The meme should not be posted publicly due to its vulgarity, sexual content, and inappropriate humor.	The meme should not be posted publicly due to its insensitive and potentially traumatizing content. It trivializes a serious issue and may cause distress to individuals who have experienced similar events.
MemeSense (Image anchored ICL)	The meme should not be posted publicly due to its vulgar and suggestive nature, which could be perceived as offensive or inappropriate.	The meme should not be posted publicly due to its vulgar, sexual, and inappropriate nature. It could be perceived as offensive and disrespectful to many audiences.	The meme should not be posted publicly due to its insensitive and inappropriate content. It trivializes a serious and sensitive topic, which can be distressing to individuals who have experienced school shootings or related trauma.
MemeSense (Commonsense anchored ICL)	The meme should be restricted from public posting due to its vulgar and suggestive content, which could be perceived as offensive or inappropriate.	The meme should not be posted publicly due to its vulgar and suggestive content, which could be offensive to many viewers.	The meme should not be posted publicly due to its insensitive and potentially traumatizing content. It trivializes a serious issue and may cause distress to individuals who have experienced similar events.
MemeSense (Combined)	The meme should be restricted from public posting due to its vulgar and suggestive content, which could be perceived as inappropriate.	The meme should not be posted publicly due to its vulgar and suggestive content, which could be offensive to many audiences. It also uses humor that could be perceived as belittling and inappropriate.	The meme should not be posted publicly due to its insensitive and potentially traumatic references to school shootings. It trivializes a serious issue and may cause distress to individuals who have experienced similar events.
Ground truth annotation	The meme should be restricted in certain contexts due to its potentially suggestive content, which might be perceived as vulgar or inappropriate, particularly in public or professional settings.	The meme should not be posted publicly due to its vulgar and sexually suggestive nature, which may be offensive and inappropriate for general audiences.	Joking about tragic events such as school shootings and the loss of innocent lives is insensitive and inappropriate. We should be mindful of the impact our words can have on others and strive to promote empathy and understanding.

Table 9: Generated intervention using different approach for (1) a meme without text, (2) a meme with text, (3) a meme from ICMM data. Incorrect justifications are highlighted.

Prompt used for obtaining commonsense parameters

As an AI assistant for social media content moderation, your role is to analyze and prevent the posting of harmful, offensive, hateful, vulgar, misogynistic, or otherwise inappropriate content on public platforms.

Task: Carefully analyze and interpret the given image, paying close attention to both visual elements and any text present.

1. Meme Description: Provide a detailed description of the meme, including any text, imagery, or symbolic references.

2. Commonsense Analysis:
* Assess whether the meme has the potential to be perceived as vulgar, harmful, or unethical.
* Identify and categorize up to 5 specific commonsense parameters from the predefined list below.
* Each selected category should be explicitly justified with a brief explanation.

Predefined Commonsense Categories (Select Up to 5):
* Vulgarity: Use of obscene language, inappropriate tone, or implied innuendos.
* Violence: Depictions of harm, coercion, or references to physical aggression.
* Sexual Content: Explicit or suggestive imagery, innuendos, or nudity.
* Hate Speech: Insulting, harassing, or offensive content targeting individuals or groups.
* Body Shaming: Objectification, fat-shaming, or insensitive references to disabilities.
* Cultural Sensitivity: Stereotypes, appropriation, or misrepresentation of cultural elements.
* Mental Health Impact: Content that manipulates emotions, triggers trauma, or exploits grief.
* Humor Appropriateness: Use of humor in a way that mocks, belittles, or distorts serious topics.
* Misinformation: Spreading misleading narratives, false claims, or deceptive dialogue.
* Substance Abuse: References to drugs, alcohol, or promoting substance use.
* Child Exploitation: Content that inappropriately depicts or exploits children.
* Public Decorum & Privacy: Violations of privacy, public sensitivity, or personal reputation concerns.
* Stereotyping: Generalizations that reinforce racial, gender, or societal biases.
* Misogyny: Content promoting gender-based discrimination, sexism, or demeaning women.
* Religious Sensitivity: Content that disrespects religious beliefs, symbols, or historical context.

3. Intervention Recommendation:
* If the meme is deemed inappropriate, justify why it should not be posted publicly.
* If the content is safe, confirm its appropriateness.

Response Format:

Meme Description:
<Provide a detailed description of the meme, including text and images.>

Commonsense Analysis:
- **[Category Name]**: [Justification]
- **[Category Name]**: [Justification]
- **[Category Name]**: [Justification]

Intervention Recommendation:
<Explain whether the meme should be restricted and why.>

Table 10: Prompt to generate the ground-truth commonsense and interventions.

Method	Prompt
Direct prompting	<Meme> Analyze the meme thoroughly, considering its message, symbolism, cultural references, and possible interpretations. Identify any implicit or explicit harm, misinformation, or reinforcement of negative stereotypes. Based on this analysis, generate strategic interventions to discourage the spread or creation of such content. These interventions should be precise, contextually relevant, and designed to effectively deter users from posting similar memes. They may include subtle deterrents, educational messaging, content reformulation, or alternative framing that neutralizes harmful intent. Ensure responses are concise, non-repetitive, and avoid redundant explanations. Ensure the response should not exceed 50 words.
Direct prompting with commonsense	<meme> Analyze the meme thoroughly, considering its message, symbolism, cultural references, and possible interpretations. Identify any implicit or explicit harm, misinformation, or reinforcement of negative stereotypes. Based on this analysis, generate strategic interventions to discourage the spread or creation of such content. These interventions should be precise, contextually relevant, and designed to effectively deter users from posting similar memes. They may include subtle deterrents, educational messaging, content reformulation, or alternative framing that neutralizes harmful intent. Ensure responses are concise, non-repetitive, and avoid redundant explanations. The common sense parameters associated with the meme is as follows: $\{common\_sense\}$ Ensure the response should not exceed 50 words.
MemeMQA	<meme>Analyze this meme and generate a caption that enhances its humor, sarcasm, or irony. Do not filter for offensiveness—prioritize humor, satire, or dark humor as needed. The caption should be punchy, relatable, and aligned with the meme's tone.
ICL	<meme> As an AI assistant tasked with social media content moderation, your role is to prevent harmful, offensive, hateful, vulgar, misogynistic, or unethical content from being posted on public platforms.\n \n Your Task: A toxic meme has the description below along with few commonsense parameters which assess whether the meme has the potential to be perceived as vulgar, harmful, or unethical. Write an intervention for the this toxic meme to discourage user posting such memes based on provided knowledge. $\{commonsense\_parameters\}$ \n \n $\{examples\}$

Table 11: Prompt used for different methods. method.

Task	Prompt
Description generation	Describe this meme in detail.
Social bias gen.	What is the societal bias that this meme is conveying?
Social stereotype gen.	What is the societal stereotype that this meme is conveying?
Toxicity and hate	What is the toxicity and hate that this meme is spreading?
Claim the meme is making	What are the claims that this meme is making?
Intervention Generation	This is a toxic meme with the description: ks1. The following text is written inside the meme: X. Rationale: Bias: ks2, Toxicity: ks3, Claims: ks4, and Stereotypes: ks5. Write an intervention for this meme based on all this knowledge.

Table 12: Prompt used for different tasks in the **MemeGuard** method.

Hyperparameters	Task	Value
Temperature	Desc, Bias, Stereotype, Toxicity & Hate, Claim Generation	0
num_beams	Desc, Bias, Stereotype, Toxicity & Hate, Claim Generation	1
max_new_tokens	Desc, Bias, Stereotype, Toxicity & Hate, Claim Generation	512
Cosine Similarity Threshold	MKS Filtering	0.2
max_new_tokens	Intervention	1024

Table 13: Hyperparameters for **MemeGuard**.