ExPO-HM: Learning to Explain-then-Detect for Hateful Meme Detection

Anonymous authors
Paper under double-blind review

Abstract

Hateful memes have emerged as a particularly challenging form of online abuse, motivating the development of automated detection systems. Most prior approaches rely on direct detection, producing only binary predictions. Such models fail to provide the context and explanations that real-world moderation requires. Recent Explain-then-Detect approaches, using Chain-of-Thought prompting or LMM agents, perform worse than simple SFT baselines, and even advanced post-training methods such as GRPO fail to close the gap. Our analysis identifies two key issues of such systems: important policy-relevant cues such as targets and attack types are not hypothesized by the model as a likely explanation; and the binary reward signal is insufficient to guide reasoning. To address these challenges, we propose ExPO-HM (Explain-then-Detect Policy Optimization for Hateful Memes), inspired by the training and evaluation process of human annotators. ExPO-HM combines SFT warmup, GRPO with curriculum learning, and Conditional Decision Entropy (CDE) as both metric and reward for reasoning quality. Across three hateful meme benchmarks, ExPO-HM achieves state-of-the-art performance on binary detection, fine-grained classification, and reasoning quality, with up to 15% and 17% F1 improvement over the GRPO and DPO baselines, respectively. By moving hateful meme detection from simple binary alarms to explanation-driven detection, ExPO-HM provides accurate, interpretable, and actionable moderation support.

This paper contains content for demonstration purposes that may be disturbing for some readers.

1 Introduction

The rise of social media has led to a surge in hateful content, notably in the form of memes. This has sparked growing research interest in automated hateful meme detection systems that aim at supporting human moderation (Kiela et al., 2021; Liu et al., 2022; Prakash et al., 2023; Shah et al., 2024). Most prior work focuses on direct detection, which only provides a binary classification as to whether a meme is hateful or benign (Cao et al., 2023; Mei et al., 2024; Su et al., 2025). However, recent studies show that moderators require additional information to improve efficiency (Calabrese et al., 2024), such as what type of attack is present, and why the system considers the meme harmful. Additionally, social media users may also benefit from understanding these explanations of harmfulness.

Interestingly, human annotators are not trained and evaluated on binary judgments; common practice is that they are guided by a detailed moderation policy manual that defines policy violations such as disparagement of protected groups (Singhal et al., 2023). It would be infeasible to train annotators by showing them only raw examples with binary labels; the fine-grained framework provides the necessary structure for both training and evaluation. This human analogy highlights a crucial gap: if humans require fine-grained guidelines and reasoning to make reliable judgments, automated systems could benefit from the same. We call this setting "Explain-then-Detect", where the system first generates a natural language rationale and then produces a classification decision.

Recent work builds Explain-then-Detect Large Multimodal Model (LMM) systems using Chain-of-Thought (CoT) prompting (Wei et al., 2023; Pan et al., 2025) or agent-based frameworks (Huang et al., 2024), but these perform worse than direct Supervised Fine-tuning (SFT) baselines (Mei et al.,

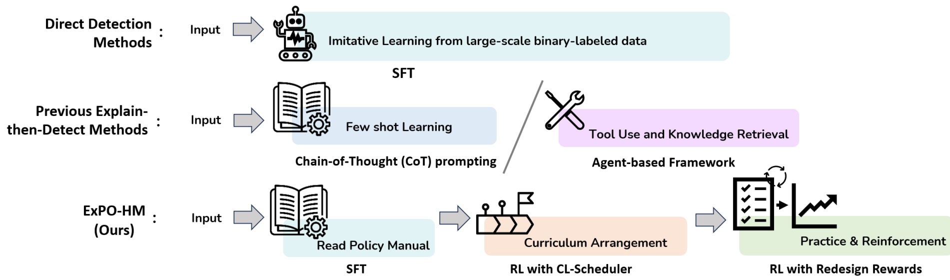

Figure 1: Comparing previous methods with ExPO-HM.

2025). Reinforcement learning methods such as Group-Relative Policy Optimization (GRPO) (Shao et al., 2024) can strengthen model reasoning through post-training, yet we find that applying GRPO directly still underperforms SFT for hateful meme detection. Our study reveals two key challenges for Explain-then-Detect systems. First, model explanations often fail to identify the correct violated policy or target, leading to misleading predictions. Second, the binary reward signal in GRPO is too weak to guide reasoning, just as human annotators cannot learn from only yes/no labels.

To address these issues, we propose ExPO-HM (**Ex**plain-then-Detect **P**olicy **O**ptimization for **H**ateful **M**emes), inspired by how human annotators are trained and evaluated. ExPO-HM first uses SFT warmup on a policy manual, mirroring the guideline-based training of human annotators. We then apply GRPO with curriculum learning, mimicking how annotators are first trained and evaluated on fine-grained categories before making binary judgments. We further introduce Conditional Decision Entropy (CDE) both as a metric for explanation quality and as a reward signal to encourage decisive reasoning. We summarize our contributions:

- **Paradigm.** We introduce the first Explain-then-Detect hateful meme detection that outperforms direct detection, enabling accurate and interpretable hateful meme understanding.
- **Methods.** ExPO-HM mimics human moderator training, combining policy manual SFT warmup, GRPO curriculum learning, and CDE-based reward optimization.
- **Evaluation.** We propose a comprehensive evaluation setup that reflects real-world moderation, extending beyond binary classification to fine-grained categories and hateful reasoning judged by LLMs, with extensive baseline comparisons.
- **Results.** ExPO-HM surpasses previous best systems, and achieves new state-of-the-art performance across binary, fine-grained, and reasoning benchmarks, with up to **15%** and **17%** F1 improvement over the GRPO and DPO baseline, respectively.

2 Related Work

Direct Hateful Meme Detection Most existing approaches to hateful meme detection treat the task as binary classification. Numerous studies fine-tune CLIP-based models using only binary labels and train dedicated classifiers (Pramanick et al., 2021; Kumar & Nandakumar, 2022; Burbi et al., 2023; Cao et al., 2023; Ji et al., 2024; Mei et al., 2024). Decoder-based LMMs have also been fine-tuned for this task (Alayrac et al., 2022; Laurençon et al., 2023; Hu et al., 2024). In particular, Mei et al. (2025) train a classifier and retriever on top of the LMM embeddings, achieving state-of-the-art binary detection performance.

In contrast, fine-grained classification, such as identifying attack types or target groups, has received far less attention, despite its importance in real-world moderation. Annotated datasets are available (Mathias et al., 2021a; Dimitrov et al., 2021; Fersini et al., 2022; Shah et al., 2024), and some earlier work has explored this problem (Zia et al., 2021; Mathias et al., 2021b), but recent progress has been limited. Mod-Hate (Cao et al., 2024) and IntMeme (Hee & Lee, 2025) leverage fine-grained annotations during training but do not report fine-grained results. MemeCLIP (Shah et al., 2024) addresses this by fine-tuning separate CLIP-based classifiers for each split. In this paper, we systematically evaluate models under different setups and extend the evaluation to fine-grained classification, addressing this important gap.

Explain-then-Detect Hateful Meme Detection Compared to direct hateful meme classification, research on explainable hateful meme detection is far more limited. With the rise of decoder-based language models, some Explain-then-Detect systems have emerged. For example, Lin et al. (2024) leverage a debate between two LMs to decide meme harmfulness, while LOREHM (Huang et al., 2024) adopts a reasoning-agent framework with retrieval and reflection. However, these systems still primarily target binary classification.

A key challenge is the lack of annotated explanation data. Hatred (Hee et al., 2023), built on the Facebook Hateful Memes dataset (Kiela et al., 2021), remains the only open-source dataset with human-written rationales. Other efforts, such as the recent Arabic hateful meme dataset ArMeme (Kmainasi et al., 2025), are not yet publicly available. Moreover, reasoning tasks remain difficult (Nguyen & Ng, 2024). Existing Explain-then-Detect systems not only struggle with reasoning but also underperform direct detection models in binary classification, underscoring the cost of requiring explanations without tailored optimization strategies. In this paper, we make two key contributions. First, we benchmark a comprehensive set of Explain-then-Detect systems using the Hatred dataset. Second, inspired by human moderator training, we develop ExPO-HM, the first Explain-then-Detect system that surpasses both prior explainable and direct detection approaches, delivering accurate and interpretable hateful meme detection.

3 ExPO-HM Methodology

3.1 Preliminaries

Problem Statement. A common binary hateful memes classification dataset (Kiela et al., 2021) is $\mathcal{D} = \{(I_i, c_i^*)\}_{i=1}^N$, where $I_i \in \mathbb{R}^{C\times H\times W}$ is an image with overlaid text, and the ground-truth label $c_i^* \in \{0, 1\}$ denotes `benign` / `hateful`. In addition, we consider annotations including fine-grained labels z_i^* (e.g., protected category, attack type) (Mathias et al., 2021a) and, when available, gold explanations (Hee et al., 2023) $\mathbf{e}_i^*$. We thus define the three tasks for hateful meme detection: (1) predicting binary class c_i; (2) predicting fine-grained class z_i; (3) generating $\mathbf{e}_i$. For text-based evaluation, we denote the textualized label prediction as d_i (from c_i or z_i) and the corresponding ground-truth text label as d_i^*.

Large Multimodal Models (LMMs). Given a meme I and a prompt p, we denote the input to LMM as $\mathbf{x} = (I, p)$. An LMM with parameters θ defines an auto-regressive policy over output text tokens $\mathbf{y} = (y_1, \ldots, y_{|\mathbf{y}|})$:

$$\pi_\theta(\mathbf{y} \mid \mathbf{x}) \;=\; \prod_{t=1}^{|\mathbf{y}|} \pi_\theta(y_t \mid y_{<t}, \mathbf{x}), \tag{1}$$

where t indexes the output tokens. *Direct-Detection* methods decode labels directly, via answers like “`yes`” / “`no`” (Lin et al., 2024). In contrast, *Explain-then-Detect* first generates reasoning and then the label. Following the standard long CoT format (DeepSeek-AI et al., 2025), the output sequence is:

$$\mathbf{y} \equiv \big(\texttt{<think>}\ \mathbf{e}\ \texttt{</think>}\ \texttt{<answer>}\ d\ \texttt{</answer>}\big), \tag{2}$$

where $\mathbf{e}$ is the generated explanation and d is the textualized label prediction.

Supervised Fine-Tuning (SFT). Given an input $\mathbf{x}$ and a target output sequence $\mathbf{y}^*$, the model is trained by maximizing the likelihood of $\mathbf{y}^*$:

$$\mathcal{L}_{\text{SFT}}(\theta) = -\sum_{t=1}^{|\mathbf{y}^*|} \log \pi_\theta(y_t^* \mid \mathbf{y}_{<t}^*, \mathbf{x})\,. \tag{3}$$

This serves as the general form of SFT used in our baselines.

Direct Preference Optimization (DPO). We consider DPO (Rafailov et al., 2023) as a baseline fine-tuning method. Preference pairs $(\mathbf{y}^+, \mathbf{y}^-)$ are sampled on-policy from the reference model π_{ref} via the Explain-then-Detect prompting format. A response $\mathbf{y}$ is selected as the preferred response $\mathbf{y}^+$ if its decision d matches the ground-truth label d^*; otherwise, it is treated as the rejected response $\mathbf{y}^-$.

We optimize the DPO objective:

$$\mathcal{L}_{\text{DPO}}(\theta) = -\log \sigma \left(\beta \log \frac{\pi_\theta(\mathbf{y}^+|\mathbf{x})}{\pi_{\text{ref}}(\mathbf{y}^+|\mathbf{x})} - \beta \log \frac{\pi_\theta(\mathbf{y}^-|\mathbf{x})}{\pi_{\text{ref}}(\mathbf{y}^-|\mathbf{x})} \right), \tag{4}$$

where σ is the sigmoid function and π_{ref} is the reference model, i.e., the initial model before DPO fine-tuning.

Group Relative Policy Optimization (GRPO). GRPO (Shao et al., 2024) is an online Policy Gradient method that discards the critic model to save computation. To estimate the advantage, it samples a group of outputs $(\mathbf{y}_1, \ldots, \mathbf{y}_G)$ from the old policy $\pi_{\theta_{\text{old}}}$ for each input $\mathbf{x}$. The advantage for the g-th sample in a group is computed by normalizing its reward against the group's reward distribution $\{r_1, \ldots, r_G\}$:

$$A_g = \frac{r_g - \text{mean}(\{r_1, \ldots, r_G\})}{\text{std}(\{r_1, \ldots, r_G\})}. \tag{5}$$

We consider verifiable reward functions in this paper. The policy is then optimized with the clipped objective:

$$\mathcal{L}_{\text{GRPO}}(\theta) = \frac{1}{G} \sum_{i=1}^{G} \left[\min \left(\frac{\pi_\theta(\mathbf{y}_i|\mathbf{x})}{\pi_{\theta_{\text{old}}}(\mathbf{y}_i|\mathbf{x})} A_i, \text{clip}\left(\frac{\pi_\theta(\mathbf{y}_i|\mathbf{x})}{\pi_{\theta_{\text{old}}}(\mathbf{y}_i|\mathbf{x})}, 1-\epsilon, 1+\epsilon \right) A_i \right) \right] - \beta D_{\text{KL}}(\pi_\theta \| \pi_{\text{ref}}). \tag{6}$$

3.2 Conditional Decision Entropy

The reasoning quality is difficult to optimize in hateful meme detection, as there is no reliable reward model due to the scarce taionale corpora and subjectuive human judgements. To address this, we propose Conditional Decision Entropy (CDE) as a proxy measure. The principle of CDE is straightforward: good reasoning should lead to a sharp and correct decision, while poor reasoning produces confusion.

CDE Definition. For an input $\mathbf{x}$, the LMM π_θ generates an explanation and decision response $\mathbf{y} = (\mathbf{e}, d) \sim \pi_\theta(\cdot \mid \mathbf{x})$ in the format of Eq. 2, where the final decision is sampled conditioned on the explanation and input $d \sim \pi_\theta(\cdot \mid \mathbf{e}, \mathbf{x})$. We define CDE as the entropy of the decision *conditioned on* the produced explanation:

$$H(d \mid \mathbf{e}, \mathbf{x}) = -\, \mathbb{E}_{d \sim \pi_\theta(\cdot|\mathbf{e},\mathbf{x})} \big[\log \pi_\theta(d \mid \mathbf{e}, \mathbf{x}) \big]. \tag{7}$$

Monte Carlo Estimator for CDE To evaluate reasoning quality with CDE, we estimate average CDE over the validation set. For each example $\mathbf{x}_i$, we sample $K = 16$ explanations $\mathbf{e}_{ik}$ with the policy π_θ and compute the entropy of the decision distribution. The estimator is

$$\widehat{H}(d \mid \mathbf{e}, \mathbf{x}) \;=\; \frac{1}{K|\mathcal{D}|} \sum_{i=1}^{|\mathcal{D}|} \sum_{k=1}^{K} H(d \mid \mathbf{e}_{ik}, \mathbf{x}_i)\,, \quad \mathbf{e}_{ik} \sim \pi_\theta(\cdot \mid \mathbf{x}_i). \tag{8}$$

In the binary classification case, we experimented with collapsing the decision vocabulary to $\mathcal{V} \in \{\texttt{Yes}, \texttt{No}\}$, making CDE equivalent to binary entropy. We observed no significant difference compared to using the full vocabulary. For generalizability to fine-grained multi-class labels, we therefore adopt the full vocabulary formulation.

A full derivation is provided in Appendix E.

3.3 ExPO-HM Framework

Inspired by human moderator training, where annotators first study policy guidelines and then practice applying them to tasks of increasing difficulty, ExPO-HM, as shown in Figure 2, first learn policy knowledge through SFT, then refines its reasoning via GRPO with curriculum learning, progressing from fine-grained to binary classification.

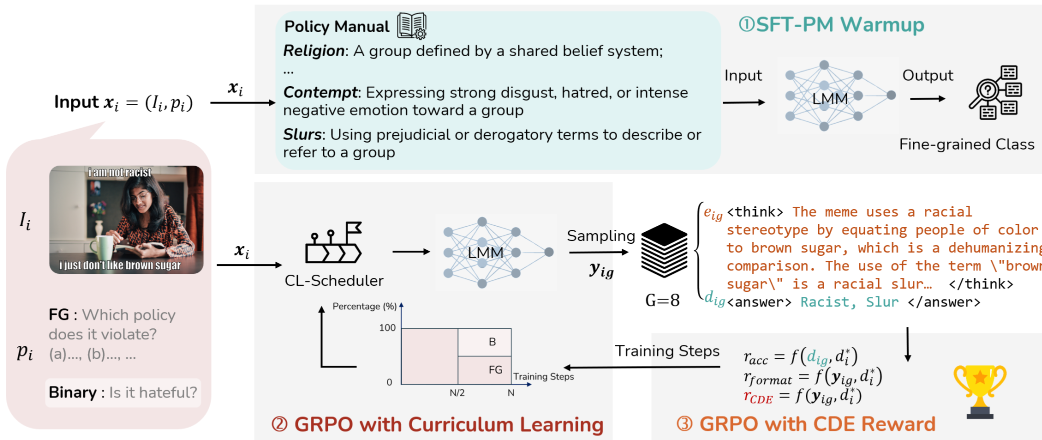

Figure 2: Architecture of ExPO-HM. Our framework consists of three key components: ① **SFT-PM Warmup.** The VLM is first trained with SFT using structured policy manuals derived from fine-grained labels and dataset guidelines, teaching the model to align decisions with explicit moderation policies. ② **GRPO with Curriculum Learning**. Training follows a two-stage schedule: the first 50% of steps use fine-grained data only for reasoning exploration, and the remaining 50% use a balanced 50/50 mix of fine-grained and binary data. ③ **GRPO with CDE reward** In addition to the format reward (r_{format}) and accuracy (r_{acc}) reward used in standard GRPO, we also add a Conditional Decision Entropy (r_{CDE}) reward.

SFT Warmup on Structured Policy Manuals (SFT-PM). We first teach the LMM moderation policy knowledge by converting each dataset's fine-grained labels into a structured policy manual as the input prompt. Descriptions derived from the dataset annotation guidelines are added to each policy item in the policy manual, where we provide details in Appendix B. We optimize the language modelling loss in Eq. 3 with this policy manual augmented input for each meme, and the target response $\mathbf{y}^*$ is the fine-grained label d_i^*. Note that we do not use human-written gold hateful explanation e^* in the warmup stage, as they are off-policy and lead to worse performance, which we discuss in Sec 4.5.

GRPO with Curriculum Learning (GRPO-CL). After the SFT-PM warmup, we conduct GRPO curriculum learning. We begin with fine-grained classification to incentivize policy understanding through diverse reasoning exploration, then introduce binary classification for hateful vs. benign detection. We test various curriculum schedulers, switching after fine-grained accuracy plateaus, adjusting the budget split between stages, or adjusting the mixing ratio of the fine-grained and binary data in the second stage, and find similar performance as long as fine-grained reasoning precedes binary. We therefore adopt a simple 50/50/50 strategy: the first 50% of steps use fine-grained data only, and the remaining 50% use a balanced 50/50 mix of fine-grained and binary data.

We optimize the clipped surrogate loss in Eq. 6 using the group-relative advantage in Eq. 5. The reward r_{ig} corresponds to the g-th response in the sampled group for the i-th training example

$$r(\mathbf{y}_{ig}, d_i^*) = r_{\text{format}} + r_{\text{acc}} + w\, r_{\text{CDE}}, \tag{9}$$

where $r_{\text{format}} \in \{0, 1\}$ checks if the output obeys the correct template in Eq. 2. The accuracy reward $r_{\text{acc}} \in [0, 1]$ measures prediction correctness with partial credit for multi-class fine-grained classification and penalties for over-prediction. For binary classification, it requires an exact match and thus $r_{\text{acc}} \in \{0, 1\}$. For the GRPO baseline, we set $w = 0$, leaving only the format and accuracy rewards. Now let's define CDE Reward r_{CDE}.

CDE as a Reward Although GRPO with curriculum learning improves over the naive GRPO baseline, it still falls short in producing reliable reasoning. As introduced in Sec. 3.2, CDE provides a proxy for reasoning quality. If the prediction is sharp and correct, the reasoning is helpful and should be rewarded; if it is wrong but confident, the reasoning is misleading and should be penalized. We therefore incorporate it as an additional reward to guide ExPO-HM.

For each group-sampled example $\mathbf{y}_{ig}$ of each input $\mathbf{x}_i$, we denote the CDE as h_{ig} and correctness as δ_{ig}:

$$h_{ig} = H(d \mid \mathbf{e}_{ig}, \mathbf{x}_i), \qquad \delta_{ig} = \mathbf{1}\big[d_{ig} = d_i^*\big]. \tag{10}$$

We reward confident correctness, tolerate uncertainty when wrong, and penalize confident errors. The CDE reward for the example $\mathbf{y_{ig}}$ is

$$r_{\mathrm{CDE}}(h_{ig}, \delta_{ig}) = \delta_{ig} \cdot \begin{cases} w, & h \leq a \\ w\,\dfrac{b - h_{ig}}{b - a}, & a < h_{ig} < b \\ 0, & h_{ig} \geq b \end{cases} + (1 - \delta_{ig}) \cdot \begin{cases} -\rho w, & h_{ig} \leq a \\ w\,\dfrac{h_{ig} - a}{b - a}, & a < h_{ig} < b \\ w, & h_{ig} \geq d \end{cases} \tag{11}$$

CDE rewards contribute a maximum of weight w, with ρ controlling the penalty strength for over-confident wrong predictions. Unless otherwise noted, we use default hyperparameters $a = 0.1$, $b = 0.5$, $w = 0.2$, and $\rho = 0.25$. The r_{CDE} can thus be fed into Eq. 9 to obtain the reward to compute advantage to optimize the GRPO objective.

4 EXPERIMENTS

4.1 EXPERIMENTAL SETUP

Dataset: We evaluate the binary and fine-grained classification on three meme classification datasets: HatefulMemes (Kiela et al., 2021), MAMI (Fersini et al., 2022), and PrideMM (Shah et al., 2024).

Tasks: We evaluate binary classification (hateful vs benign) on all three datasets. For fine-grained classification, we assess attack methods and target groups on HatefulMemes, attack methods on MAMI, and stance towards LGBTQ+ along with target group detection on PrideMM. Due to the scarcity of annotated hate rationales, we only evaluate reasoning quality on HatefulMemes, where gold human rationales are available (Hee et al., 2023). Detailed dataset descriptions and statistics are provided in Appendix B.

Evaluation Metrics: We evaluate classification tasks using macro F1 following prior work (Shah et al., 2024). For reasoning quality, we adopt the LLM-as-judge method (Yang et al., 2023; Mei et al., 2025) to measure alignment between model-generated and human rationales. In addition, we report CDE as a proxy to reasoning quality and verify its correlation with LLM-as-judge in Appendix D.

4.2 BASELINES

We compare ExPO-HM with comprehensive baselines on Qwen2.5-VL-3B and Qwen2.5-VL-7B (Bai et al., 2025) in Table 1. In this section, we describe the baseline setup briefly. Full implementation details are provided in Appendix C to ensure reproducibility.

SFT. In this paper, we consider two variants of SFT as baselines. *Direct-SFT* is trained with the ground-truth label as the target ($\mathbf{y}^* = d^*$ in Eq. 3), while *CoT-SFT* uses Explain-then-Detect prompt adopted in DPO and GRPO, where the target sequence is the chosen response in DPO sampling ($\mathbf{y}^* = \mathbf{y}^+$ in Eq. 3). In practice, we find that *Direct-SFT* consistently outperforms *CoT-SFT*, even when inference is performed with the Explain-then-Detect prompt. We therefore report *Direct-SFT* as the default baseline. For classification, we train and report separate models based on the binary and the fine-grained subset, and report the best results. Full results for each model are provided in Table 3, while Table 1 reports the best system.

DPO & GRPO. For DPO and GRPO, we initialize from the fine-grained SFT warmup, but without the policy-manual style augmentation. We sweep different β values in DPO to get the best performance on the validation set. The GRPO baseline is trained with the same compute budget as ExPO-HM, using identical hyperparameter settings in both the warmup and GRPO fine-tuning stages.

Best prior systems. We compare ExPO-HM with the best prior systems. RA-HMD (Mei et al., 2025) is the state-of-the-art direct detection model, combining two-stage fine-tuning, and retrieval-augmented classification. Although primarily designed for direct detection, it supports reasoning evaluation via prompting, so we report its LLM-as-judge scores. All RA-HMD results are based on Qwen2.5-VL-7B (Bai et al., 2025). For Explain-then-Detect, we compare two recent systems: LOREHM (Huang et al., 2024), a reflective reasoning agent with tool-calling capability built on LLaVA-Next-34B (Liu et al., 2024), and U-CoT+ (Pan et al., 2025), which uses human-guided CoT prompting with Qwen2-VL-7B (Wang et al., 2024) for meme-to-text conversion and Qwen2.5-14B (Qwen et al., 2025) for answer generation. We can only report their results on the binary classification due to their prompt-based agent design, these systems cannot be directly adapted for fine-grained classification or structured reasoning tasks. Furthermore, we did not include closed-source reasoning LMMs such as the OpenAI o-series (OpenAI, 2024) as baselines, since over 30% of requests were blocked by the API server due to the harmful nature of the examples.

Table 1: Comparing ExPO-HM with baseline systems across three datasets. *B* stands for Binary and *R* stands for Reasoning. LLM refer to the LLM-as-judge score. Best results are in **bold**. ↑ indicates higher is better, ↓ lower is better.

#	Model	HatefulMemes					MAMI			PrideMM			
		B	*Attack*	*Target*	*R*		*B*	*Attack*	*R*	*B*	*Stance*	*Target*	*R*
		F1 ↑	F1 ↑	F1 ↑	LLM ↑	CDE ↓	F1 ↑	F1 ↑	CDE ↓	F1 ↑	F1 ↑	F1 ↑	CDE ↓
	Direct Detection Baselines												
1	Qwen2.5-VL-3B												
2	*Zero-shot*	53.1	42.1	60.1	-	-	61.1	48.2	-	58.6	53.7	48.8	-
3	*SFT*	71.9	64.3	69.3	-	-	77.9	61.8	-	74.3	58.6	53.2	-
4	Qwen2.5-VL-7B												
5	*Zero-shot*	59.8	50.3	60.2	-	-	63.4	50.2	-	65.2	56.8	51.1	-
6	*SFT*	75.0	64.7	71.1	-	-	78.1	63.1	-	75.6	60.2	61.0	-
7	*RA-HMD*	80.2	-	-	5.4	-	81.0	-	-	77.8	-	-	-
	Explain-then-Detect Systems												
8	LOREHM (34B)	65.6	-	-	-	-	75.3	-	-	-	-	-	-
9	U-CoT+ (14B)	72.4	-	-	-	-	79.9	-	-	71.4	-	-	-
10	Qwen2.5-VL-3B												
11	*Zero-shot*	52.5	41.7	58.7	3.3	0.42	58.7	41.7	0.32	52.6	51.2	40.8	0.33
12	*SFT*	62.3	62.7	63.3	3.6	0.40	69.2	60.1	0.34	63.2	56.6	49.8	0.29
13	*DPO*	59.6	52.3	58.1	3.5	0.42	66.8	50.2	0.36	64.2	55.5	48.9	0.34
14	*GRPO*	63.4	55.6	66.1	3.8	0.32	76.6	61.2	0.19	72.1	57.3	48.4	0.18
15	*ExPO-HM*	74.7	71.5	73.7	5.1	0.16	80.7	70.4	0.08	75.6	66.5	62.1	0.12
16	Qwen2.5-VL-7B												
17	*Zero-shot*	65.9	44.7	64.5	5.0	0.33	63.9	46.5	0.23	59.4	54.6	50.2	0.28
18	*SFT*	74.5	58.4	69.4	5.0	0.33	72.8	62.6	0.19	68.3	58.0	50.9	0.28
19	*DPO*	73.6	63.2	66.6	4.9	0.32	72.3	56.6	0.22	69.5	56.3	52.3	0.30
20	*GRPO*	74.5	61.2	64.5	5.2	0.26	76.8	63.7	0.09	73.2	58.6	60.1	0.14
21	*ExPO-HM*	**81.1**	**75.6**	**77.2**	**6.2**	**0.03**	**82.3**	**73.0**	**0.04**	**78.7**	**68.4**	**65.1**	**0.08**

4.3 Comparing ExPO-HM to Baseline Systems

Table 1 compares ExPO-HM with the aforementioned baseline post-training methods and state-of-the-art systems. We report qualitative examples in Appendix E.4. Here, we summarize the key observations.

Baseline Explain-then-Detect methods hurt classification performance. Under *Explain-then-Detect*, post-training variants (SFT/DPO/GRPO, #18-#20 for Qwen2.5-VL-7B) consistently underperform the Direct-Detection SFT baseline (#6), except for the comparable performance on the MAMI Attack classification. Larger agentic and CoT systems (LOREHM, U-CoT+, #8-#9) also fall short of strong Direct-Detection baselines like SFT and RA-HMD (#6-#7). For instance, the binary classification on HatefulMemes is 80.2 on RA-HMD vs 72.4 F1 with U-CoT+. On HatefulMemes binary classification, RA-HMD reaches 80.2 F1, compared to 72.4 with U-CoT+. Explain-then-Detect systems are crucial for building automatic moderation systems that can truly support real-world moderators, but these results highlight that simply adding explicit rationales through CoT prompting or standard post-training hurts classification accuracy. This motivates the design of ExPO-HM, which aims to improve Explain-then-Detect systems without sacrificing predictive performance.

Naive post-training barely improves performance. Explain-then-Detect post-training (#18–#20) improves classification over zero-shot (#17), but reasoning quality stagnates, failing to meet the goal of improving reasoning through post-training. On HatefulMemes with Qwen2.5-VL-7B, the zero-shot LLM-as-judge score is 5.0; DPO drops below this, while GRPO only nudges it to 5.2. Even with online RL, reasoning remains difficult to improve. Moreover, post-training still underperforms strong CoT systems specifically designed for hateful meme detection (#8-#9). This underscores the need for dedicated post-training methods like ExPO-HM, which are tailored to hateful meme detection and designed to improve not only classification accuracy but also the quality of explanations.

ExPO-HM consistently outperforms. ExPO-HM delivers the strongest performance across binary detection, fine-grained classification, and reasoning. On Qwen2.5-VL-7B, it surpasses RA-HMD and all post-training baselines, achieving large gains in fine-grained F1 (**+14.4** on HatefulMemes Attack, **+12.7** on Target, compared to GRPO with equal compute). Reasoning also improves markedly, with 6.2 on LLM-as-judge vs. 5.2 for GRPO. These results confirm ExPO-HM's effectiveness across datasets and tasks.

Strong correlation between LLM-as-judge metric and CDE metric. On the HatefulMemes reasoning dataset, we observe a strong alignment between the LLM-as-judge score and the CDE score. To quantify this, we evaluate the correlation based on results from all the reported setups, with three random seeds each, yielding 60 data points. We find a strong negative correlation (Pearson $r = -0.78$, Spearman $\rho = -0.81$, both $p < 0.001$), confirming that lower CDE values, reflecting more confident and accurate reasoning, correspond to higher reasoning quality.

4.4 Ablation Study of ExPO-HM Components

We conduct an ablation study to examine the contribution of the three key components in ExPO-HM. Results on HatefulMemes with Qwen2.5-VL-7B are reported in Table 2. Without SFT-PM, the warmup falls back to SFT with fine-grained labels without policy manual augmentation. Without GRPO-CL, GRPO is trained on a randomly mixed set of binary and fine-grained data. Without CDE, GRPO uses only the format and accuracy rewards.

Table 2: Ablation Study of ExPO-HM Components.

#	Components			HatefulMemes				
	SFT-PM	GRPO-CL	CDE	*B* F1 ↑	*Attack* F1 ↑	*Target* F1 ↑	*R* LLM ↑	CDE ↓
1	-	-	-	74.5	61.2	64.5	5.2	0.263
2	✓	-	-	75.8	70.8	70.2	5.6	0.092
3	✓	✓	-	78.4	74.3	76.1	5.8	0.056
4	✓	✓	✓	81.1	75.6	77.2	6.2	0.026

SFT-PM enhanced the fine-grained warmup. Compared to the baseline warmup without policy manual augmentation (#1), SFT-PM improves performance across all metrics. This indicates that fine-grained labels alone are insufficient for policy understanding, while policy manual augmentation substantially strengthens both classification and reasoning. In Sec. 4.5, we further present a systematic comparison of different warmup strategies.

GRPO-CL further improves performance. Building on SFT-PM, adding curriculum learning to GRPO (#3) yields further gains across the board. The key difference is ordering, GRPO-CL first let the model explore on reasoning over the fine-grained labels before binary classification. This order proves crucial: standard GRPO produces short average responses (28 tokens) in binary classification, while GRPO-CL nearly doubles this (52 tokens), indicating not only higher quality but also more detailed reasoning is incentivized during training.

CDE improves both accuracy and explanation quality. Adding CDE on top of SFT-PM and GRPO-CL further immproves the perforamnce. Noteably, LLM-judge score improved to 6.2, and a marked drop in CDE 0.026, suggesting that the model's rationales become more aligned with sharp, correct decisions.

Table 3: Comparing SFT warmup variants on HatefulMemes on Qwen2.5-VL-7B: no warmup (-), SFT on binary labels (SFT-B), SFT on gold reasoning (SFT-R), SFT on fine-grained labels (SFT-FG), and SFT with policy-manual augmentation (SFT-PM).

#	Warmup	SFT				w/ GRPO-CL and CDE			
		B F1 ↑	*Attack* F1 ↑	*Target* F1 ↑	*R* LLM ↑	*B* F1 ↑	*Attack* F1 ↑	*Target* F1 ↑	*R* LLM ↑
1	-	65.9	44.7	64.5	5.0	73.3	69.3	72.1	5.2
2	SFT-B	74.1	58.2	69.4	4.9	73.5	66.8	70.1	5.1
3	SFT-R	72.2	51.6	63.1	5.0	79.2	72.3	73.2	5.7
4	SFT-FG	72.5	58.4	67.7	4.9	78.9	73.4	73.4	5.6
5	SFT-PM	74.3	64.6	68.8	5.0	81.1	75.6	77.2	6.2

4.5 Effects of different Warmup Strategy

Table 3 compares five warmup strategies for Qwen2.5-VL-7B on the HatefulMemes dataset. For each, we report the Explain-then-Detect performance after SFT and the performance after GRPO-CL with CDE reward .

Good SFT does not necessarily transfer to good RL performance. Although SFT-B performs better than SFT-R and SFT-FG at the SFT stage, its performance after GRPO-CL is comparably worse than its counterparts, even below the no-warmup baseline. This suggests that binary-only warmup fails to equip the model with the moderation concepts needed for reasoning-guided RL. In contrast, our proposed SFT-PM explicitly teaches such concepts via policy manual augmentation, yielding both stronger SFT performance and the best results after ExPO-HM training.

4.6 CDE Analysis

Figure 3 presents box-and-whisker plots of CDE distributions for ExPO-HM and GRPO on the HatefulMemes validation set with Qwen2.5-VL-7B. ExPO-HM maintains very low CDE for correct predictions ($\mu = 0.019$) and higher CDE for wrong ones ($\mu = 0.048$), yielding a clear separation. In contrast, the GRPO baseline shows high CDE for both correct ($\mu = 0.278$) and wrong predictions ($\mu = 0.226$), showing weaker separation. This demonstrates that ExPO-HM produces reasoning that is not only more accurate but also better aligned with decision confidence.

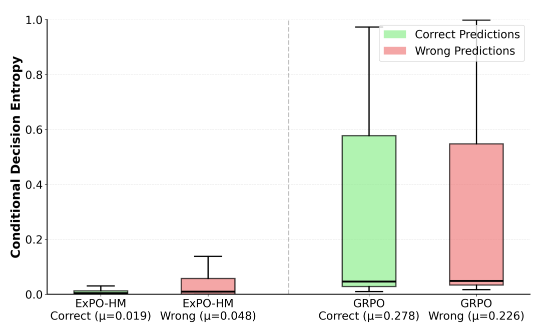

Figure 3: Comparison of CDE distributions between ExPO-HM and GRPO on the HatefulMemes validation set with Qwen2.5-VL-7B.

We also test whether adding the CDE reward causes policy entropy collapse, a phenomenon reported in prior RL work (Cui et al., 2025) when entropy bonuses are removed. Our results show that overall policy entropy remains comparable to the baseline GRPO system without CDE, confirming that the CDE reward, acting only on the decision part of the generation, does not reduce exploration.

5 Conclusion

We propose ExPO-HM, which combines SFT warmup on policy-manual–augmented data with GRPO curriculum learning, guided by a Conditional Decision Entropy reward to promote high-quality reasoning. Comprehensive experiments show that ExPO-HM achieves state-of-the-art performance on binary detection, fine-grained classification, and reasoning quality.

ETHICAL STATEMENT

Societal benefits. Hateful meme detection systems such as ExPO-HM can help automatically identify and mitigate harmful online content, reducing the prevalence of hate speech. By providing explanations in addition to predictions, our system not only supports safer digital environments for end-users but also alleviates the burden on human content moderators, improving their wellbeing. We believe such systems play an essential role in fostering respectful online communication and contributing to healthier digital communities.

Intended use. We will enforce strict access controls for releasing model checkpoints and artifacts. Access will be limited to researchers who agree to our terms of use, which explicitly restrict the system to detection and prevention of hateful speech. Any use that promotes, condones, or encourages hate speech or other harmful content is strictly prohibited.

Misuse potential. Although ExPO-HM is not designed to introduce bias, it is trained on datasets that may reflect societal or annotator biases (Pramanick et al., 2021). These biases could propagate into model predictions. To mitigate risks of unfair or disproportionate moderation, human oversight remains essential when deploying such systems.

Deployment considerations. Moderation of hateful content is inherently influenced by cultural norms and subjective judgments. Expressions considered benign in one context may be offensive in another. Since ExPO-HM is trained with policy manuals, its outputs depend critically on the underlying moderation policies. Careful review and adaptation of community guidelines are therefore crucial to ensure responsible deployment across diverse cultural and linguistic contexts.

Usage of Datasets. The datasets used in this study, HatefulMemes, MAMI, and PrideMM, were curated for research purposes to combat online hate speech. We strictly adhere to the terms of use established by the dataset authors.

REPRODUCIBILITY STATEMENT

We provide detailed dataset usage and statistics in Appendix 4, including all augmentation procedures introduced in this paper. For the evaluation of reasoning, we detail the process in Appendix D. Experimental setups, implementation details including code base usage, software version, hardware configuration, and hyperparameter settings are described in Appendix C to facilitate reproducibility. Upon publication, we will release the source code on GitHub and make all artifacts, including augmented data, sampled training sets, and trained checkpoints available to HuggingFace.

REFERENCES

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob L. Menick, Sebastian Borgeaud, Andy Brock, Aida Nematzadeh, Sahand Sharifzadeh, Mikołaj Bińkowski, Ricardo Barreira, Oriol Vinyals, Andrew Zisserman, and Karén Simonyan. Flamingo: a visual language model for few-shot learning. *Advances in Neural Information Processing Systems*, 35: 23716–23736, Dec 2022. URL `https://openreview.net/forum?id=EbMuimAbPbs`.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-vl technical report. (arXiv:2502.13923), February 2025. doi: 10.48550/arXiv.2502.13923. URL `http://arxiv.org/abs/2502.13923`. arXiv:2502.13923 [cs].

Giovanni Burbi, Alberto Baldrati, Lorenzo Agnolucci, Marco Bertini, and Alberto Del Bimbo. Mapping memes to words for multimodal hateful meme classification. In *2023 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, pp. 2824–2828, 2023. doi: 10.1109/ICCVW60793.2023.00303.

Agostina Calabrese, Leonardo Neves, Neil Shah, Maarten W. Bos, Björn Ross, Mirella Lapata, and Francesco Barbieri. Explainability and hate speech: Structured explanations make social media moderators faster. (arXiv:2406.04106), 2024. doi: 10.48550/arXiv.2406.04106. URL http://arxiv.org/abs/2406.04106. arXiv:2406.04106 [cs].

Rui Cao, Ming Shan Hee, Adriel Kuek, Wen-Haw Chong, Roy Ka-Wei Lee, and Jing Jiang. Pro-cap: Leveraging a frozen vision-language model for hateful meme detection. In *Proceedings of the 31st ACM International Conference on Multimedia*, MM '23, pp. 5244–5252. Association for Computing Machinery, 2023. ISBN 9798400701085. doi: 10.1145/3581783.3612498. URL https://doi.org/10.1145/3581783.3612498.

Rui Cao, Roy Ka-Wei Lee, and Jing Jiang. Modularized networks for few-shot hateful meme detection. In *Proceedings of the ACM Web Conference 2024*, WWW '24, pp. 4575–4584. Association for Computing Machinery, 2024. ISBN 9798400701719. doi: 10.1145/3589334.3648145. URL https://doi.org/10.1145/3589334.3648145.

Ganqu Cui, Yuchen Zhang, Jiacheng Chen, Lifan Yuan, Zhi Wang, Yuxin Zuo, Haozhan Li, Yuchen Fan, Huayu Chen, Weize Chen, Zhiyuan Liu, Hao Peng, Lei Bai, Wanli Ouyang, Yu Cheng, Bowen Zhou, and Ning Ding. The entropy mechanism of reinforcement learning for reasoning language models, May 2025. URL https://arxiv.org/abs/2505.22617v1.

DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, Aixin Liu, Bing Xue, Bingxuan Wang, Bochao Wu, Bei Feng, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, Damai Dai, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, Fuli Luo, Guangbo Hao, Guanting Chen, Guowei Li, H. Zhang, Han Bao, Hanwei Xu, Haocheng Wang, Honghui Ding, Huajian Xin, Huazuo Gao, Hui Qu, Hui Li, Jianzhong Guo, Jiashi Li, Jiawei Wang, Jingchang Chen, Jingyang Yuan, Junjie Qiu, Junlong Li, J. L. Cai, Jiaqi Ni, Jian Liang, Jin Chen, Kai Dong, Kai Hu, Kaige Gao, Kang Guan, Kexin Huang, Kuai Yu, Lean Wang, Lecong Zhang, Liang Zhao, Litong Wang, Liyue Zhang, Lei Xu, Leyi Xia, Mingchuan Zhang, Minghua Zhang, Minghui Tang, Meng Li, Miaojun Wang, Mingming Li, Ning Tian, Panpan Huang, Peng Zhang, Qiancheng Wang, Qinyu Chen, Qiushi Du, Ruiqi Ge, Ruisong Zhang, Ruizhe Pan, Runji Wang, R. J. Chen, R. L. Jin, Ruyi Chen, Shanghao Lu, Shangyan Zhou, Shanhuang Chen, Shengfeng Ye, Shiyu Wang, Shuiping Yu, Shunfeng Zhou, Shuting Pan, S. S. Li, Shuang Zhou, Shaoqing Wu, Shengfeng Ye, Tao Yun, Tian Pei, Tianyu Sun, T. Wang, Wangding Zeng, Wanjia Zhao, Wen Liu, Wenfeng Liang, Wenjun Gao, Wenqin Yu, Wentao Zhang, W. L. Xiao, Wei An, Xiaodong Liu, Xiaohan Wang, Xiaokang Chen, Xiaotao Nie, Xin Cheng, Xin Liu, Xin Xie, Xingchao Liu, Xinyu Yang, Xinyuan Li, Xuecheng Su, Xuheng Lin, X. Q. Li, Xiangyue Jin, Xiaojin Shen, Xiaosha Chen, Xiaowen Sun, Xiaoxiang Wang, Xinnan Song, Xinyi Zhou, Xianzu Wang, Xinxia Shan, Y. K. Li, Y. Q. Wang, Y. X. Wei, Yang Zhang, Yanhong Xu, Yao Li, Yao Zhao, Yaofeng Sun, Yaohui Wang, Yi Yu, Yichao Zhang, Yifan Shi, Yiliang Xiong, Ying He, Yishi Piao, Yisong Wang, Yixuan Tan, Yiyang Ma, Yiyuan Liu, Yongqiang Guo, Yuan Ou, Yuduan Wang, Yue Gong, Yuheng Zou, Yujia He, Yunfan Xiong, Yuxiang Luo, Yuxiang You, Yuxuan Liu, Yuyang Zhou, Y. X. Zhu, Yanhong Xu, Yanping Huang, Yaohui Li, Yi Zheng, Yuchen Zhu, Yunxian Ma, Ying Tang, Yukun Zha, Yuting Yan, Z. Z. Ren, Zehui Ren, Zhangli Sha, Zhe Fu, Zhean Xu, Zhenda Xie, Zhengyan Zhang, Zhewen Hao, Zhicheng Ma, Zhigang Yan, Zhiyu Wu, Zihui Gu, Zijia Zhu, Zijun Liu, Zilin Li, Ziwei Xie, Ziyang Song, Zizheng Pan, Zhen Huang, Zhipeng Xu, Zhongyu Zhang, and Zhen Zhang. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. (arXiv:2501.12948), January 2025. doi: 10.48550/arXiv.2501.12948. URL http://arxiv.org/abs/2501.12948. arXiv:2501.12948 [cs].

Dimitar Dimitrov, Bishr Bin Ali, Shaden Shaar, Firoj Alam, Fabrizio Silvestri, Hamed Firooz, Preslav Nakov, and Giovanni Da San Martino. Detecting propaganda techniques in memes. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 6603–6617, Online, Aug 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.516. URL https://aclanthology.org/2021.acl-long.516.

Elisabetta Fersini, Francesca Gasparini, Giulia Rizzi, Aurora Saibene, Berta Chulvi, Paolo Rosso, Alyssa Lees, and Jeffrey Sorensen. Semeval-2022 task 5: Multimedia automatic misogyny identification. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, pp. 533–549, Seattle, United States, Jul 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.semeval-1.74. URL https://aclanthology.org/2022.semeval-1.74.

Ming Shan Hee and Roy Ka-Wei Lee. Demystifying hateful content: Leveraging large multimodal models for hateful meme detection with explainable decisions. 2025. doi: 10.48550/ARXIV.2502.11073. URL https://arxiv.org/abs/2502.11073.

Ming Shan Hee, Wen-Haw Chong, and Roy Ka-Wei Lee. Decoding the underlying meaning of multimodal hateful memes. (arXiv:2305.17678), Jun 2023. URL http://arxiv.org/abs/2305.17678. arXiv:2305.17678 [cs].

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=nZeVKeeFYf9.

Yushi Hu, Otilia Stretcu, Chun-Ta Lu, Krishnamurthy Viswanathan, Kenji Hata, Enming Luo, Ranjay Krishna, and Ariel Fuxman. Visual program distillation: Distilling tools and programmatic reasoning into vision-language models. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9590–9601, Seattle, WA, USA, June 2024. IEEE. ISBN 979-8-3503-5300-6. doi: 10.1109/CVPR52733.2024.00916. URL https://ieeexplore.ieee.org/document/10655837/.

Jianzhao Huang, Hongzhan Lin, Liu Ziyan, Ziyang Luo, Guang Chen, and Jing Ma. Towards low-resource harmful meme detection with lmm agents. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 2269–2293, Miami, Florida, USA, November 2024. Association for Computational Linguistics. URL https://aclanthology.org/2024.emnlp-main.136.

Junhui Ji, Xuanrui Lin, and Usman Naseem. Capalign: Improving cross modal alignment via informative captioning for harmful meme detection. In *Proceedings of the ACM Web Conference 2024*, pp. 4585–4594, Singapore Singapore, May 2024. ACM. ISBN 979-8-4007-0171-9. doi: 10.1145/3589334.3648146. URL https://dl.acm.org/doi/10.1145/3589334.3648146.

Douwe Kiela, Hamed Firooz, Aravind Mohan, Vedanuj Goswami, Amanpreet Singh, Pratik Ringshia, and Davide Testuggine. The hateful memes challenge: Detecting hate speech in multimodal memes. (arXiv:2005.04790), Apr 2021. URL http://arxiv.org/abs/2005.04790. arXiv:2005.04790 [cs].

Mohamed Bayan Kmainasi, Abul Hasnat, Md Arid Hasan, Ali Ezzat Shahroor, and Firoj Alam. Memeintel: Explainable detection of propagandistic and hateful memes. (arXiv:2502.16612), February 2025. doi: 10.48550/arXiv.2502.16612. URL http://arxiv.org/abs/2502.16612. arXiv:2502.16612 [cs].

Gokul Karthik Kumar and Karthik Nandakumar. Hate-CLIPper: Multimodal hateful meme classification based on cross-modal interaction of CLIP features. In *Proceedings of the Second Workshop on NLP for Positive Impact (NLP4PI)*, pp. 171–183, Abu Dhabi, United Arab Emirates (Hybrid), December 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.nlp4pi-1.20. URL https://aclanthology.org/2022.nlp4pi-1.20.

Hugo Laurençon, Lucile Saulnier, Léo Tronchon, Stas Bekman, Amanpreet Singh, Anton Lozhkov, Thomas Wang, Siddharth Karamcheti, Alexander M. Rush, Douwe Kiela, Matthieu Cord, and Victor Sanh. Obelics: An open web-scale filtered dataset of interleaved image-text documents. (arXiv:2306.16527), August 2023. URL http://arxiv.org/abs/2306.16527. arXiv:2306.16527 [cs].

Hongzhan Lin, Ziyang Luo, Wei Gao, Jing Ma, Bo Wang, and Ruichao Yang. Towards explainable harmful meme detection through multimodal debate between large language models. In *Proceedings of the ACM Web Conference 2024*, WWW '24, pp. 2359–2370, New York, NY, USA, 2024.

Association for Computing Machinery. ISBN 9798400701719. doi: 10.1145/3589334.3645381. URL https://doi.org/10.1145/3589334.3645381.

Chen Liu, Gregor Geigle, Robin Krebs, and Iryna Gurevych. Figmemes: A dataset for figurative language identification in politically-opinionated memes. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 7069–7086, Abu Dhabi, United Arab Emirates, Dec 2022. Association for Computational Linguistics. URL https://aclanthology.org/2022.emnlp-main.476.

Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. Llava-next: Improved reasoning, ocr, and world knowledge, January 2024. URL https://llava-vl.github.io/blog/2024-01-30-llava-next/.

Lambert Mathias, Shaoliang Nie, Aida Mostafazadeh Davani, Douwe Kiela, Vinodkumar Prabhakaran, Bertie Vidgen, and Zeerak Waseem. Findings of the woah 5 shared task on fine grained hateful memes detection. *Proceedings of the 5th Workshop on Online Abuse and Harms (WOAH 2021)*, pp. 201–206, 2021a. doi: 10.18653/v1/2021.woah-1.21.

Lambert Mathias, Shaoliang Nie, Aida Mostafazadeh Davani, Douwe Kiela, Vinodkumar Prabhakaran, Bertie Vidgen, and Zeerak Waseem. Findings of the WOAH 5 shared task on fine grained hateful memes detection. In Aida Mostafazadeh Davani, Douwe Kiela, Mathias Lambert, Bertie Vidgen, Vinodkumar Prabhakaran, and Zeerak Waseem (eds.), *Proceedings of the 5th Workshop on Online Abuse and Harms (WOAH 2021)*, pp. 201–206, Online, August 2021b. Association for Computational Linguistics. doi: 10.18653/v1/2021.woah-1.21. URL https://aclanthology.org/2021.woah-1.21/.

Jingbiao Mei, Jinghong Chen, Weizhe Lin, Bill Byrne, and Marcus Tomalin. Improving hateful meme detection through retrieval-guided contrastive learning. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 5333–5347, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.291. URL https://aclanthology.org/2024.acl-long.291.

Jingbiao Mei, Jinghong Chen, Guangyu Yang, Weizhe Lin, and Bill Byrne. Robust adaptation of large multimodal models for retrieval augmented hateful meme detection. (arXiv:2502.13061), 2025. doi: 10.48550/arXiv.2502.13061. URL http://arxiv.org/abs/2502.13061. arXiv:2502.13061 [cs].

Khoi P. N. Nguyen and Vincent Ng. Computational meme understanding: A survey. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 21251–21267, Miami, Florida, USA, November 2024. Association for Computational Linguistics. URL https://aclanthology.org/2024.emnlp-main.1184.

OpenAI. Openai o1 system card. (arXiv:2412.16720), December 2024. doi: 10.48550/arXiv.2412.16720. URL http://arxiv.org/abs/2412.16720. arXiv:2412.16720 [cs].

Fengjun Pan, Anh Tuan Luu, and Xiaobao Wu. Detecting harmful memes with decoupled understanding and guided cot reasoning. (arXiv:2506.08477), 2025. doi: 10.48550/arXiv.2506.08477. URL http://arxiv.org/abs/2506.08477. arXiv:2506.08477 [cs].

Nirmalendu Prakash, Ming Shan Hee, and Roy Ka-Wei Lee. Totaldefmeme: A multi-attribute meme dataset on total defence in singapore. In *Proceedings of the 14th Conference on ACM Multimedia Systems*, MMSys '23, pp. 369–375, New York, NY, USA, Jun 2023. Association for Computing Machinery. ISBN 9798400701481. doi: 10.1145/3587819.3592545. URL https://dl.acm.org/doi/10.1145/3587819.3592545.

Shraman Pramanick, Shivam Sharma, Dimitar Dimitrov, Md. Shad Akhtar, Preslav Nakov, and Tanmoy Chakraborty. Momenta: A multimodal framework for detecting harmful memes and their targets. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pp. 4439–4455, Punta Cana, Dominican Republic, Nov 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.findings-emnlp.379. URL https://aclanthology.org/2021.findings-emnlp.379.

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. Qwen2.5 technical report. (arXiv:2412.15115), January 2025. doi: 10.48550/arXiv.2412.15115. URL http://arxiv.org/abs/2412.15115. arXiv:2412.15115 [cs].

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=HPuSIXJaa9.

Siddhant Bikram Shah, Shuvam Shiwakoti, Maheep Chaudhary, and Haohan Wang. Memeclip: Leveraging clip representations for multimodal meme classification. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 17320–17332, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.959. URL https://aclanthology.org/2024.emnlp-main.959/.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. (arXiv:2402.03300), April 2024. doi: 10.48550/arXiv.2402.03300. URL http://arxiv.org/abs/2402.03300. arXiv:2402.03300 [cs].

Mohit Singhal, Chen Ling, Pujan Paudel, Poojitha Thota, Nihal Kumarswamy, Gianluca Stringhini, and Shirin Nilizadeh. Sok: Content moderation in social media, from guidelines to enforcement, and research to practice. pp. 868–895, 2023. doi: 10.1109/EuroSP57164.2023.00056. URL https://ieeexplore.ieee.org/document/10190527/.

Xuanyu Su, Yansong Li, Diana Inkpen, and Nathalie Japkowicz. A context-aware contrastive learning framework for hateful meme detection and segmentation. In Luis Chiruzzo, Alan Ritter, and Lu Wang (eds.), *Findings of the Association for Computational Linguistics: NAACL 2025*, pp. 5201–5215, Albuquerque, New Mexico, April 2025. Association for Computational Linguistics. ISBN 979-8-89176-195-7. doi: 10.18653/v1/2025.findings-naacl.289. URL https://aclanthology.org/2025.findings-naacl.289/.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. (arXiv:2409.12191), October 2024. doi: 10.48550/arXiv.2409.12191. URL http://arxiv.org/abs/2409.12191. arXiv:2409.12191 [cs].

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. (arXiv:2201.11903), Jan 2023. doi: 10.48550/arXiv.2201.11903. URL http://arxiv.org/abs/2201.11903. arXiv:2201.11903 [cs].

Yongjin Yang, Joonkee Kim, Yujin Kim, Namgyu Ho, James Thorne, and Se-Young Yun. Hare: Explainable hate speech detection with step-by-step reasoning. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 5490–5505, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-emnlp.365. URL https://aclanthology.org/2023.findings-emnlp.365/.

Haris Bin Zia, Ignacio Castro, and Gareth Tyson. Racist or sexist meme? classifying memes beyond hateful. In Aida Mostafazadeh Davani, Douwe Kiela, Mathias Lambert, Bertie Vidgen, Vinodkumar Prabhakaran, and Zeerak Waseem (eds.), *Proceedings of the 5th Workshop on Online Abuse and Harms (WOAH 2021)*, pp. 215–219, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.woah-1.23. URL https://aclanthology.org/2021.woah-1.23/.

A Use of Large Language Models

Our coding work was assisted by Github Copilot. OpenAI ChatGPT was only used in proofreading and spell-checking. We claim that the content presented in this paper was fully original.

B Dataset details and statistics

HatefulMemes (Kiela et al., 2021) Released by Meta in 2020, HatefulMemes contains 12,000 memes annotated as hateful or benign by trained experts. This benchmark dataset synthesizes memes targeting religion, race, disability, and gender. It includes confounder examples where the benign memes are generated by altering either the image or text to challenge models' ability in multimodal reasoning.

In 2021, Meta released an extended annotation set, the Fine-Grained Hateful Meme Classification dataset (Mathias et al., 2021a). This dataset augments the original HatefulMemes with detailed labels for both attack types and target groups, enabling fine-grained classification.

For policy-manual SFT, we convert these fine-grained labels into structured moderation rules:

> Does this meme use any of the following types of attack against a group?
> - Dehumanizing: Presenting a group as subhuman, explicitly or implicitly
> - Inferiority: Claiming that a group is inferior, less worthy, or less important
> - Inciting violence: Calling for or encouraging harm—physical or otherwise—towards a group
> - Mocking: Belittling or making jokes about a group
> - Contempt: Expressing strong disgust, hatred, or intense negative emotion
> - Slurs: Using prejudicial or derogatory terms to describe a group
> - Exclusion: Advocating for removal, segregation, or marginalization of a group
>
> If the meme does not use any of these attack types, respond with Benign.

> Does this meme express hate or negative bias toward any of the following protected groups?
> - Religion: A group defined by shared belief systems
> - Race: A group defined by racialized physical characteristics
> - Sex: A group defined by sexual attributes or sexual identification
> - Nationality: A group defined by country or region of origin
> - Disability: A group defined by conditions leading to permanent dependencies
>
> If the meme does not target any protected group, respond with Benign.

For fine-grained classification, we adopt the following prompts for training and evaluation:

> Does this meme express hate or negative bias toward any of the following protected groups?
> Religion, Race, Sex, Nationality, Disability
> Please respond with one or more of the protected categories if applicable. If the meme does not contain any hateful content, respond with Benign.

> Does this meme use any of the following types of attack against a group?
> Dehumanizing, Inferiority, Inciting violence, Mocking, Contempt, Slurs, Exclusion
> Please respond with one or more of the attack types if applicable. If the meme does not contain any hateful content, respond with Benign.

MAMI (Fersini et al., 2022) The MAMI dataset focuses on detecting misogynistic memes sourced from various social media platforms, including Twitter and Reddit, as well as meme creation and sharing websites, and even anti-women websites and forums. It contains annotation for two tasks: (1) binary classification of misogyny and (2) categorization of misogyny types. In this work, we use the binary task to detect whether a meme is misogynistic and treat the type categorization as a fine-grained multi-class, multi-label classification problem. Each sample can take one or more of four attack type labels: objectification, shaming, stereotype, and violence, or Benign if no harm is present.

For policy-manual SFT, we convert these fine-grained labels into structured moderation rules:

> Based on the content and context of this meme, does it use any of the following types of attack against a group?
> Choose from the following options:
> - objectification: The content reduces individuals or groups to objects, ignoring their personhood or agency
> - shaming: The content ridicules, mocks, or publicly humiliates individuals or groups
> - stereotype: The content attributes oversimplified, generalized, or exaggerated traits to individuals or groups
> - violence: The content depicts or encourages physical harm, threats, or violent actions against individuals or groups
> If the meme does not contain any hateful content, respond with Benign.
> What is the attack type?

For rest training and evaluation on fine-grained class:

> Does this meme use any of the following types of attack against a group?
> objectification, shaming, stereotype, violence
> Please respond with one or more of the attack types if applicable.
> If the meme does not contain any hateful content, respond with Benign.

PrideMM (Shah et al., 2024) PrideMM contains LGBTQ+-themed memes annotated for four tasks: hate speech detection, hate target identification, topical stance classification, and humor detection. In this work, we use the hate speech classification annotations for the binary hateful meme detection.

We further use the hate target identification and topical stance classification in our fine-grained classification setup. Both tasks are formulated as multi-class, single-label classification.

For the policy-manual SFT, we convert the fine-grained labels into structured moderation rules. For the hate target identification task, the prompt enumerates the four categories—undirected, individual, community, organization—with descriptions derived from the dataset guidelines. For the stance classification task, the policy manual lists the three categories, neutral, support, oppose, with corresponding definitions and examples.

Target identification prompt for SFT-PM:

> Based on the content and context of this meme, Does this meme express hate or negative bias toward any of the following targets?
>
> Choose from the following options:
> - undirected: General targeting without specific individuals or groups
> - individual: Targeting specific individuals
> - community: Targeting LGBTQ+ communities or groups
> - organization: Targeting specific organizations or institutions
> If the meme does not contain any hateful content, respond with Benign.
>
> What is the target type?

Stance classification prompt for SFT-PM:

> Based on the content and context of this meme, what is the stance towards LGBTQ+ individuals or communities?
>
> Choose from the following options:
> - neutral: The content does not express clear support or opposition
> - support: The content expresses positive attitudes or support
> - oppose: The content expresses negative attitudes or opposition
>
> What is the stance?

For all other fine-grained training and evaluation tasks, Target identification prompt:

> Based on the content and context of this meme, Does this meme express hate or negative bias toward any of the following targets?
>
> Choose from the following options: undirected individual community organization
>
> If the meme does not contain any hateful content, respond with Benign.
> What is the target type?

Stance classification prompt:

> Based on the content and context of this meme, what is the stance towards LGBTQ+ individuals or communities?
>
> Choose from the following options: neutral, support, oppose
>
> What is the stance?

B.1 Binary Classification statistics

Table 4 shows the data split for our binary evaluation datasets. For hateful meme, we use the `dev_seen` split as the validation set, `test_seen` as the test set.

Datasets	Train		Test	
	#Benign	#Hate	#Benign	#Hate
HatefulMemes	5450	3050	500	500
MAMI	4500	4500	500	500
PrideMM	2581	2482	260	247

Table 4: Statistical summary of binary classification datasets.

B.2 Fine-grained Classification statistics

Table 5 reports the detailed distribution of fine-grained attributes in the HatefulMemes dataset, covering both attack types and protected categories. Note that we use the `dev_unseen` split for final evaluation.

Table 6 provides the fine-grained label distributions for the MAMI dataset, focusing on Sub-task B (Type of Misogyny).

Table 7 summarizes the fine-grained label distributions for the PrideMM dataset, including both target categories and stance annotations across the training and test splits.

B.3 Hateful Reasoning Corpus Statistics

Table 8 presents the dataset statistics for Haterd, which only includes hateful memes paired with explanations. The test set corresponds to the original `HatefulMemes dev_seen` split.

Fine-grained types		**train**	**dev_unseen**	**dev_seen**
Attack type	dehumanizing	1318	104	121
	inferiority	658	35	49
	inciting_violence	407	23	26
	mocking	378	29	35
	contempt	235	6	10
	slurs	205	4	6
	exclusion	114	8	13
Protected category	religion	1078	77	95
	race	1008	63	78
	sex	746	46	56
	nationality	325	20	26
	disability	255	17	22

Table 5: Statistics of fine-grained attributes in the HatefulMemes dataset, showing attack types and protected categories across train, dev_unseen, and dev_seen splits.

Category	**Training Set**	**Test Set**
Shaming	1274 (25.48%)	146 (29.20%)
Stereotype	2810 (56.20%)	350 (70.00%)
Objectification	2202 (44.04%)	348 (69.60%)
Violence	953 (19.06%)	153 (30.60%)

Table 6: Statistics of Sub-task B in the MAMI dataset: type of misogyny labels across training and test sets. We treat this as a multilabel, multiclass fine-grained classification task.

B.4 Dataset Licenses

To access the Facebook HatefulMemes dataset, one must follow the license from Facebook[1]. HarMeme and Harm-P are distributed for research purposes only, without a license for commercial use. MAMI is under Apache License 2.0. There is no specified license for PrideMM.

C Experiment Setup and Implementation Details

SoftWare Environment. `PyTorch 2.5.1`, `CUDA 12.4`, Huggingface `Transformer 4.45.0` and `Python 3.10.12` were used for implementing the experiments. All the reported metrics were computed by `TorchMetrics 1.0.1`.

Hardware Environment. We conducted our GRPO and ExPO-HM experiments on a server equipped with 8 Nvidia H100 with 80GB of VRAM. For the DPO and SFT baselines, we use 1 GPU.

Training Details We freeze the vision module throughout fine-tuning, following the standard LMM fine-tuning protocol. We conduct all training with LoRA (Hu et al., 2022), with LoRA rank=64, $\alpha = 128$. For DPO sampling and all the inference, we use vLLM inference engine `0.9.2`.

C.1 SFT and DPO Training

For Qwen2.5-VL fine-tuning, we employ the officially recommended fine-tuning library `LLaMA-Factory 0.9.3`[2] with official hyperparameter settings for all training tasks in both the SFT and DPO, except for the LoRA config that we mentioned above. For DPO, we sweep $\beta = 0.1, 0.3, 0.5, 0.7, 0.9$ and report the best results. For all runs, we train for 3 epochs, and then select the best checkpoint based on validation performance.

[1] https://hatefulmemeschallenge.com/#download
[2] https://github.com/hiyouga/LLaMA-Factory

Fine-grained types		**Train**	**Test**
Target	Benign	2208	260
	Undirected	666	68
	Individual	219	19
	Community	986	122
	Organization	249	38
Stance	Neutral	1252	140
	Support	1645	182
	Oppose	1431	185

Table 7: Statistics of fine-grained attributes in the PrideMM dataset, showing target categories and stance labels across training and test sets.

	train	**test**	**total**
#Haterd (hateful memes only)	2,982	246	3,228

Table 8: Statistics of the Haterd dataset, which only includes hateful memes with explanations. The test set corresponds to the original `HatefulMemes dev_seen` split.

C.2 GRPO TRAINING

We use verl library `verl 0.4.1` [3]. We use the default hyperparameter settings for all training except for the LoRA configuration. For all runs, we train for 3 epochs, and then select the best checkpoint based on validation performance. The run time for ExPO-HM is about 4 hours on 8 GPUs, which is the same for the baseline GRPO experiment.

D EVALUATION OF MODEL GENERATED REASONING

Following prior work (Yang et al., 2023; Mei et al., 2025), we assess explanation quality using an LLM judge. Specifically, we provide GPT-4o-mini (`gpt-4o-mini-2024-07-18`) with reference explanations from (Hee et al., 2023). Following previous works, we adopt the same prompt as previous work:

> Compare the model-generated reasoning with the reference human reasoning for this hateful meme.
>
> Reference: {reference_reasoning}
> Model: {model_reasoning}
> Model Prediction: {model_prediction}
>
> Rate how well the model reasoning aligns with the reference on a scale of 0-10:
> - 9-10: Excellent alignment, captures all key points
> - 7-8: Good alignment, captures most key points
> - 5-6: Satisfactory alignment, captures some key points
> - 3-4: Poor alignment, misses many key points
> - 1-2: Very poor alignment, minimal understanding
> - 0: Completely wrong or unrelated
>
> Score: [0-10]
>
> Explanation: [1-2 sentences]

[3] https://github.com/volcengine/verl/releases

E CDE Derivation and Alternative estimator

E.1 CDE Derivation

Here, we provide the full derivation for the CDE metrics estimation. We consider the CDE metrics for a model parameter with θ over a dataset $\mathcal{D}$.

$$
\begin{aligned}
H(d \mid \mathbf{e}, \mathbf{x}) &= \mathbb{E}_{\mathbf{x}\sim\mathcal{D},\, \mathbf{e}\sim\pi_\theta(\mathbf{x}),\, d\sim\pi_\theta(\mathbf{e},\mathbf{x})} \left[-\log p_\theta(d \mid \mathbf{e}, \mathbf{x})\right] \\
&\quad \text{By Monte Carlo over the dataset} \\
&\approx -\frac{1}{|\mathcal{D}|}\sum_{\mathbf{x}\in\mathcal{D}} \mathbb{E}_{\mathbf{e}\sim\pi_\theta(\mathbf{x}),\, d\sim\pi_\theta(\mathbf{e},\mathbf{x})} \left[\log p_\theta(d \mid \mathbf{e}, \mathbf{x})\right] \\
&\quad \text{By Monte Carlo sampling } K \text{ times} \\
&= -\frac{1}{|\mathcal{D}|}\sum_{\mathbf{x}\in\mathcal{D}} \sum_{i=1}^{K} p_\theta(\mathbf{e}_i \mid \mathbf{x}) \sum_{d} p_\theta(d \mid \mathbf{e}_i, \mathbf{x}) \log p_\theta(d \mid \mathbf{e}_i, \mathbf{x}) \\
&\quad \text{Approximate } p_\theta(\mathbf{e}_i \mid \mathbf{x}) \approx \tfrac{1}{K} \\
&\approx -\frac{1}{K|\mathcal{D}|}\sum_{\mathbf{x}\in\mathcal{D}} \sum_{i=1}^{K} \underbrace{\sum_{d} p_\theta(d \mid \mathbf{e}_i, \mathbf{x}) \log p_\theta(d \mid \mathbf{e}_i, \mathbf{x})}_{H(d \mid \mathbf{e}_i, \mathbf{x})}
\end{aligned}
\tag{12}
$$

For the entropy $H(d \mid \mathbf{e}_i, \mathbf{x})$, we by default compute it over the full decision vocabulary:

$$
H(d \mid \mathbf{e}_i, \mathbf{x}) = -\sum_{d\in\mathcal{V}} p_\theta(d \mid \mathbf{e}_i, \mathbf{x}) \log p_\theta(d \mid \mathbf{e}_i, \mathbf{x}),
\tag{13}
$$

where $\mathcal{V}$ denotes the output vocabulary. For practical efficiency, we do not compute entropy over the entire vocabulary; instead, we approximate it using the top 10–50 tokens by likelihood, which substantially reduces computation and memory costs. When a fine-grained class is represented by multiple tokens, we compute the average token entropy similar to the policy entropy computation.

For binary classification, one may collapse the vocabulary into `Yes`, `No` by grouping all tokens semantically aligned with "yes/positive" or "no/negative," and normalizing their probabilities.

E.2 Alternative Estimator through Chain Rule

When considering the CDE, we can expand through: By the chain rule of entropy:

$$
\underbrace{H((\mathbf{e}, d) \mid \mathbf{x})}_{(1)} = \underbrace{H(\mathbf{e} \mid \mathbf{x})}_{(2)} + \underbrace{H(d \mid \mathbf{e}, \mathbf{x})}_{(3)}
\tag{14}
$$

1. $H((\mathbf{e}, d) \mid \mathbf{x})$ Sequence entropy: the total entropy of generating both reasoning and decision.
2. $H(\mathbf{e} \mid \mathbf{x})$ Reasoning entropy: measures the diversity of reasoning paths the model can produce for an input.
3. $H(d \mid \mathbf{e}, \mathbf{x})$ Conditional decision entropy (CDE): quantifies the uncertainty of the model's decision given its own on-policy reasoning path $\mathbf{e}$.

Both (1) and (2) can be estimated directly via sequence-level sampling. Then CDE, (3) can be obtained by subtraction, using the chain rule in Eq. 14.

E.3 Alternative Estimator when logits is not available

When model logits are not accessible, we approximate entropy by sampling $K = 16$ responses directly from the LMM and measuring entropy over the final detection decisions.

$$H(d \mid \mathbf{x}) \;=\; -\sum_{d \in \texttt{Yes},\texttt{No}} p_\theta(d \mid \mathbf{x}), \log p_\theta(d \mid \mathbf{x}), \tag{15}$$

where $p_\theta(d \mid \mathbf{x})$ is estimated by counting the relative frequencies of positive vs. negative decisions among the K sampled responses. This can similarly be applied towards fine-grained classes.

To approximate full CDE without logits, one can fix a reasoning trajectory and resample K responses $\mathbf{y}_{ik}$ conditioned on that reasoning (e.g., by sampling with temperature 0.7–1.0). The Monte Carlo estimator in Eq. 12 is then applied to obtain the CDE for each sampled reasoning path. Finally, averaging across multiple such sampled reasonings provides an overall CDE estimate.

E.4 Comparing GRPO and ExPO-HM Predictions

Table 9 and Table 10 show examples where ExPO-HM corrects prediction errors made by the GRPO baseline on Qwen2.5-VL-7B, illustrating its ability to generate higher-quality explanations.

	Case 1	Case 2	Case 3
Meme			
Ground Truth GRPO Prediction ExPO-HM Prediction	#Hateful #Benign #Hateful	#Hateful #Benign #Hateful	#Hateful #Benign #Hateful
GRPO Explanation	The meme is a humorous observation about a sandwich maker that can be used as an ironing board. It does not contain any offensive or derogatory content directed at any specific group or individual. The joke is based on a practical use of a kitchen appliance and does not express any form of hate or discrimination.	The meme uses a visual pun by showing dogs and labeling them as ä group of men.T̈his is a common internet meme format that uses animals to represent a group of people, often for humorous effect. The image itself does not contain any explicit content or language that could be considered hateful. The use of animals to represent a group of people is a common meme trope and is generally not intended to be offensive.	The meme uses a wheelchair as a visual metaphor for the ḧardest part to eat in a vegetable.T̈his is a play on words and is not intended to be offensive. The joke is based on a common phrase and the visual of a wheelchair, which is a tool for mobility, not a derogatory symbol. The meme is not targeting any specific group or race, and the humor is based on a general observation rather than any harmful stereotype.”,
ExPO-HM Explanation	The meme makes a derogatory comparison by using a woman’s body to represent an ironing board, which is a dehumanizing and objectifying portrayal. This type of humor is often used to degrade women and is considered dehumanizing.	The meme uses a derogatory term to refer to a group of men, which is a form of dehumanizing and can be considered degrading. The use of a dog to represent a group of men is also an example of speciesism, which is a form of prejudice against non-human animals. While the meme is not explicitly violent, it does contain elements that can be seen as dehumanizing and degrading	The meme uses a derogatory term v̈egetableẗo refer to a person in a wheelchair, which is a form of dehumanizing and derogatory language. This type of language is considered hateful as it dehumanizes and degrades the person in the wheelchair

Table 9: Visualization of cases from GRPO Qwen2.5-VL-7B and ExPO-HM Qwen2.5VL-7B Models.

	Case 4	Case 5	Case 6
Meme			
Ground Truth	#Hateful	#Hateful	#Hateful
GRPO Prediction	#Benign	#Benign	#Benign
ExPO-HM Prediction	#Hateful	#Hateful	#Hateful
GRPO Explanation	The meme depicts a person covered in soap suds, which is a common reaction to too much soap in a dishwasher. The text humorously suggests that someone has added too much soap to the dishwasher again. There are no elements in the image or text that indicate any form of discrimination or hate towards a specific group. The meme is a light-hearted joke about a common household issue.	The meme makes a claim about the existence of Muhammad based on a scientific discovery, which is a religious and historical topic. However, the meme does not use any offensive language or imagery that would be considered hate speech. It is a humorous and speculative statement, not a direct attack on any group or individual.	The meme appears to be a humorous take on a political situation, using exaggerated expressions and a juxtaposition of images to convey a message. The top image shows a person with a bearded face, possibly representing a political figure. The bottom image shows a person aiming a shotgun, which could be interpreted as a metaphor for taking back what was claimed. The use of such imagery and text is not inherently hateful but rather a form of political commentary or satire.
ExPO-HM Explanation	The meme uses a derogatory term d̈ishwasherïn a way that could be seen as dehumanizing, as it reduces a person to a machine. This type of language can be considered dehumanizing and is often used in a derogatory manner.	The meme makes a false and absurd claim about the existence of Muhammad based on a fictional scenario involving human DNA in a goat carcass. This is not a real scientific discovery and is intended to be humorous at the expense of religious beliefs. The content is not respectful to any religious group and can be seen as mocking or degrading.",	The meme contains a provocative statement and imagery that could be interpreted as threatening or aggressive towards muslims. The use of a religious figure and the phrase ẅe're taking it backs̈uggests a sense of reclaiming or asserting dominance, which can be seen as a form of hate speech.

Table 10: Part 2 of the Visualization.