They Said Memes Were Harmless — We Found the Ones That Hurt: Decoding Jokes, Symbols, and Cultural References

Sahil Tripathi
sahilkrtr@gmail.com
Jamia Hamdard
New Delhi, India

Gautam Siddharth Kashyap
gautam.kashyap@hdr.mq.edu.au
Macquarie University
Sydney, New South Wales, Australia

Mehwish Nasim
mehwish.nasim@uwa.edu.au
The University of Western Australia
Perth, Australia

Jian Yang
jian.yang@mq.edu.au
Macquarie University
Sydney, New South Wales, Australia

Jiechao Gao*
jiechao@stanford.edu
Stanford University
California, United States

Usman Naseem*
usman.naseem@mq.edu.au
Macquarie University
Sydney, New South Wales, Australia

Abstract

Meme-based social abuse detection is challenging because harmful intent often relies on implicit cultural symbolism and subtle cross-modal incongruence. Prior approaches, from fusion-based methods to in-context learning with Large Vision-Language Models (LVLMs), have made progress but remain limited by three factors: i) cultural blindness (missing symbolic context), ii) boundary ambiguity (satire vs. abuse confusion), and iii) lack of interpretability (opaque model reasoning). We introduce CROSS-ALIGN+, a three-stage framework that systematically addresses these limitations: (1) *Stage I* mitigates cultural blindness by enriching multimodal representations with structured knowledge from ConceptNet, Wikidata, and Hatebase; (2) *Stage II* reduces boundary ambiguity through parameter-efficient LoRA adapters that sharpen decision boundaries; and (3) *Stage III* enhances interpretability by generating cascaded explanations. Extensive experiments on five benchmarks and eight LVLMs demonstrate that CROSS-ALIGN+ consistently outperforms state-of-the-art methods, achieving up to 17% relative F1 improvement while providing interpretable justifications for each decision.

CCS Concepts

• **Computing methodologies** → **Machine learning approaches**; *Computer vision tasks*; *Natural language processing*; • **Information systems** → Content analysis and understanding; Social networks.

Keywords

Meme abuse detection, multimodal learning, large vision-language models, cultural knowledge grounding, interpretability, contrastive fine-tuning

ACM Reference Format:
Sahil Tripathi, Gautam Siddharth Kashyap, Mehwish Nasim, Jian Yang, Jiechao Gao, and Usman Naseem. 2026. They Said Memes Were Harmless — We Found the Ones That Hurt: Decoding Jokes, Symbols, and Cultural References. In *Proceedings of the ACM Web Conference 2026 (WWW '26), April 13–17, 2026, Dubai, United Arab Emirates.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3774904.3792718

*Corresponding Author

Figure 1: Toy Example of Implicit Abuse in Memes. Illustrative samples showing how memes encode harmful intent through subtle cross-modal incongruence and cultural symbolism. Although humorous in isolation, cultural context exposes exclusionary or abusive meanings—e.g., the alt-right appropriation of "Pepe the Frog", gendered subservience, or racially charged imagery—highlighting the need for culturally grounded multimodal reasoning in CROSS-ALIGN+.

1 Introduction

Memes have become one of the most viral forms of communication on the web, shaping online discourse on a wide range of platforms including X (Twitter), Reddit, and TikTok [16, 17]. Although often humorous, memes are increasingly weaponized to propagate hate, exclusion, and disinformation (see Figure 1) [10, 16], exploiting their ability to encode harmful intent through subtle cross-modal cues. This makes meme-based abuse detection a critical challenge for web-scale content moderation systems [11, 12]. Detecting abusive memes requires reasoning about *implicit cultural symbolism* and *cross-modal incongruence* [10, 16]. For example, a meme featuring Pepe the Frog with the caption "Welcome to our neighborhood" appears harmless in isolation, but cultural context reveals Pepe's co-optation as an alt-right symbol, transforming the message into

WWW '26, Dubai, United Arab Emirates

ACM ISBN 979-8-4007-2307-0/2026/04
https://doi.org/10.1145/3774904.3792718

exclusionary propaganda [14]. Current systems fail because they cannot ground such visual-textual signals within broader cultural knowledge [2].

Recent advances in Large Vision-Language Models (LVLMs) [3, 13, 25] offer strong multimodal understanding, yet remain limited in meme abuse detection due to: (i) cultural blindness —missing symbolic context and cultural references [17, 18]; (ii) boundary ambiguity —confusing satire with abuse [1, 24]; and (iii) lack of interpretability —providing opaque predictions without explanations [4, 26]. Previous approaches from fusion-based methods (e.g. DISARM [18], MemeCLIP [17]) to in-context learning strategies [4, 26], have improved performance, but do not address these limitations systematically [12, 23].

Our Approach. We propose CROSS-ALIGN+ (*Cross-modal Social Signal Alignment with External Knowledge & Adaptive Contrastive Reasoning*), a three-stage framework that directly mitigates the core limitations of existing LVLM-based methods. **Stage I**(*External Knowledge Grounding*), enriches multimodal representations with structured cultural knowledge from ConceptNet [19], Wikidata [21], and Hatebase [6], thereby addressing cultural blindness. **Stage II**(*Adaptive Contrastive Fine-Tuning*), leverages parameter-efficient LoRA adapters [9] with contrastive objectives to sharpen decision boundaries and reduce boundary ambiguity between satire and abuse. **Stage III**(*Hierarchical Multistage Reasoning*), introduces cascaded natural language explanations that enhance interpretability and trustworthiness [22, 26]. In summary, this work makes three main contributions:

- We present the first framework that systematically integrates external cultural knowledge into LVLM-based meme abuse detection.
- We design a three-stage framework that explicitly aligns each stage with a core limitation: cultural blindness (Stage I), boundary ambiguity (Stage II), and lack of interpretability (Stage III).
- We conduct extensive empirical analysis across five benchmarks and eight LVLMs, showing up to 17% relative F1 improvements and strong gains on culturally grounded abuse patterns.

2 Related Works

Our work is based on two key lines of research: i) detection of multimodal abuse with LVLMs, and ii) approaches to knowledge integration and efficient model adaptation. We position CROSS-ALIGN+ within these areas to highlight the unique gaps it addresses.

2.1 Multimodal Abuse Detection and LVLMs

Early multimodal abuse detection methods relied on feature fusion of visual and textual signals [10], but such approaches struggle with implicit abuse where harmful intent emerges from subtle cross-modal incongruence [1]. More recent models leverage pretrained LVLMs, such as MemeCLIP [17], and specialized designs like DISARM [18] for victim-aware detection or invariant learning for robustness [24]. While these advances improve performance, they lack the ability to ground interpretations in cultural context, a key factor for understanding symbolic abuse (e.g., political or historical symbols). In-context learning has been explored as a lightweight way to adapt LVLMs without parameter updates [22]. Extensions such as Hummingbird [4] and Critic-V [26] show promise for retrieval-augmented prompting and error detection. However, these approaches remain constrained by semantic similarity, often missing cases where abusive meaning depends on culturally loaded symbolism. CROSS-ALIGN+ directly addresses this gap by systematically linking LVLM representations to structured cultural knowledge and aligning decisions with symbolic reasoning.

2.2 Knowledge Integration and Efficient Adaptation

Knowledge integration has proven to be effective in NLP tasks such as factual reasoning [15]. For multimodal understanding, LEMMA [23] recently demonstrated that external knowledge can enhance misinformation detection. Yet, systematic integration of cultural knowledge for abuse detection remains underexplored. Resources such as ConceptNet [19], Wikidata [21], and Hatebase [6] provide complementary symbolic knowledge, but prior works have not combined them in a principled pipeline for meme abuse detection. CROSS-ALIGN+ fills this gap by using these resources to contextualize visual-textual signals with cultural grounding. In parallel, parameter-efficient fine-tuning methods such as adapters [8] and LoRA [9] enable model adaptation without prohibitive compute. These techniques have been adopted for multimodal tasks but have rarely been explored in the context of contrastive alignment for abuse detection. CROSS-ALIGN+ extends this line by coupling LoRA-based contrastive fine-tuning with knowledge-grounded demonstrations, improving decision boundaries while preserving LVLM generalization.

3 Methodology

3.1 Problem Formulation

We consider meme-based social abuse detection as a binary classification problem augmented with external cultural knowledge. Each meme is a pair (I, T) with an image $I \in \mathbb{R}^{H\times W\times 3}$ and an overlaid/associated text T. The goal is to predict $\hat{y} \in \{0, 1\}$ indicating *abusive* (1) vs. *benign* (0) content as shown in Equation (1), where Θ are model parameters and $\mathcal{K}$ denotes external cultural knowledge.

$$\hat{y} = f_\Theta(I, T, \mathcal{K}); \qquad f_\Theta : (\mathbb{R}^{H\times W\times 3}, \mathcal{V}^*) \times \mathcal{K} \to \{0, 1\}, \quad (1)$$

Why knowledge is required. Abusive meaning often depends on *cross-modal incongruence* and *culturally loaded symbols*. Thus, f_Θ should depend not only on (I, T) but also on knowledge items $K \subset \mathcal{K}$ that contextualize entities in (I, T) as shown in Equation (2), where $\mathcal{E}(I, T)$ extracts entities from both modalities, ϕ links entities to KB entries, and Retrieve gathers culturally relevant facts/symbols.

$$\hat{y} = f_\Theta\big(I, T, K(I, T)\big), \quad K(I, T) = \text{Retrieve}\big(\phi(\mathcal{E}(I, T))\big), \quad (2)$$

Learning. Given a training set $\mathcal{D} = \{(I_i, T_i, y_i)\}_{i=1}^N$, we learn Θ by minimizing a classification objective, and—crucially—aligning representations with a contrastive objective that separates abusive/benign cases under knowledge context as shown in Equation (3), where $\hat{p}_i = \sigma(g_\Theta(I_i, T_i, K_i))$ are model probabilities and $\lambda > 0$

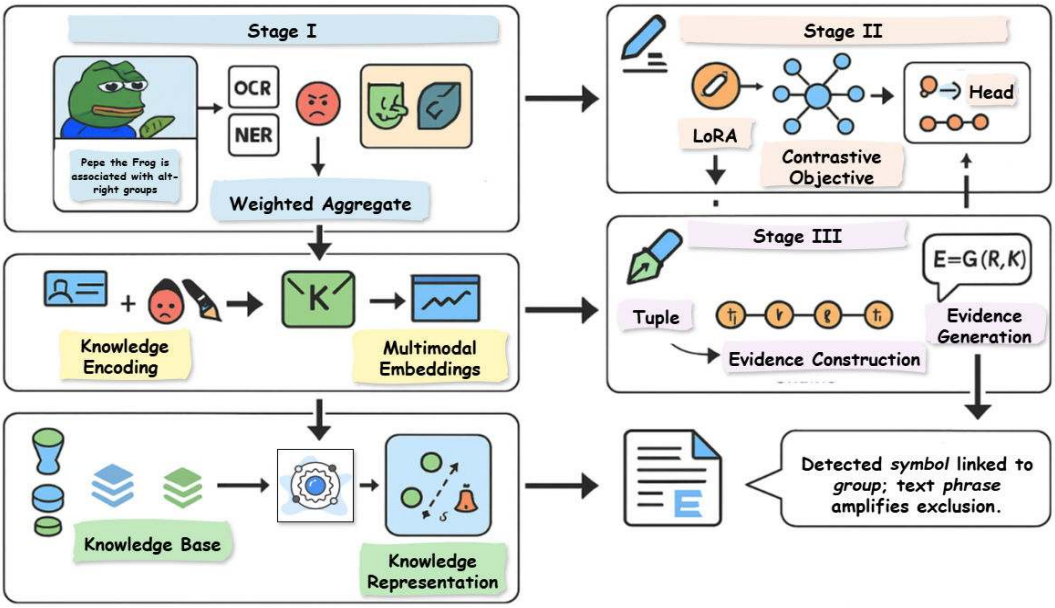

Figure 2: Overview of CROSS-ALIGN+. The pipeline consists of three stages: (I) external knowledge grounding using ConceptNet, Wikidata, and Hatebase to enrich multimodal representations; (II) adaptive contrastive fine-tuning with LoRA to improve boundary discrimination; and (III) hierarchical multistage reasoning to generate interpretable explanations. CROSS-ALIGN+ outputs both abuse predictions and faithful justifications for transparent meme moderation.

balances the losses. Section 3.2 details the construction of K_i and the contrastive term.

$$\mathcal{L} = \mathcal{L}_{\text{cls}} + \lambda\, \mathcal{L}_{\text{contrast}}, \qquad \mathcal{L}_{\text{cls}} = -\frac{1}{N}\sum_{i=1}^{N}\left[y_i \log \hat{p}_i + (1-y_i)\log(1-\hat{p}_i)\right], \tag{3}$$

3.2 CROSS-ALIGN+

CROSS-ALIGN+ (see Figure 2) is a three-stage framework aligned *one-to-one* with the limitations identified in Section 1: (I) cultural blindness, (II) boundary ambiguity, and (III) lack of interpretability. Each stage receives the output from the previous one, transforms that with stage-specific objectives, and passes enriched representations forward. Let LVLM_θ denote a pre-trained LVLM encoder producing a multimodal representation $\mathbf{h}_{\text{mm}}$ from (I, T).

3.2.1 Stage I: External Knowledge Grounding. Stage I mitigates *cultural blindness* by grounding multimodal inputs in structured cultural knowledge. Its output is a knowledge-enriched representation $\tilde{\mathbf{h}}$ that is passed to Stage II.

Entity Extraction and Linking. We extract entities from both modalities: $\mathcal{E}_I = \text{OCR/OD}(I)$, $\mathcal{E}_T = \text{NER}(T)$, $\mathcal{E} = \mathcal{E}_I \cup \mathcal{E}_T$. Entities are linked via $\phi : \mathcal{E} \to \mathcal{K}$ to entries in ConceptNet [19], Wikidata [21], and Hatebase [6]. For each $e \in \mathcal{E}$ we retrieve source-specific knowledge $\{k_{e,c}, k_{e,w}, k_{e,h}\}$ and construct a weighted aggregate: $K = \sum_{e\in\mathcal{E}} \left(w_c\, k_{e,c} + w_w\, k_{e,w} + w_h\, k_{e,h}\right)$.

Knowledge Encoding and Fusion. We encode K using the text encoder native to the chosen LVLM (e.g., if DeepSeek-VL-S is used, its own text encoder performs the encoding), which processes linearized knowledge triples (e.g., "Pepe the Frog is associated with alt-right groups") into embeddings. This yields $\mathbf{M}_K \in \mathbb{R}^{L_K \times d}$ aligned with the multimodal representation $\mathbf{h}_{\text{mm}}$. Given $\mathbf{h}_{\text{mm}} \in \mathbb{R}^{L\times d}$ from $\text{LVLM}_\theta(I, T)$, we fuse knowledge using cross-attention as shown in Equation (4), where $\tilde{\mathbf{h}}$ is the knowledge-grounded representation. A lightweight *gate* can optionally modulate the residual: $\tilde{\mathbf{h}} = \mathbf{h}_{\text{mm}} + \sigma\big(\mathbf{W}_g[\mathbf{h}_{\text{mm}}; \text{pool}(\mathbf{M}_K)]\big) \odot \text{CrossAttn}(\cdot)$. This $\tilde{\mathbf{h}}$ is the input to Stage II.

$$\tilde{\mathbf{h}} = \mathbf{h}_{\text{mm}} + \text{CrossAttn}\big(Q = \mathbf{h}_{\text{mm}}, K = \mathbf{M}_K, V = \mathbf{M}_K\big), \tag{4}$$

3.2.2 Stage II: Adaptive Contrastive Fine-Tuning. Stage II mitigates *boundary ambiguity* by refining $\tilde{\mathbf{h}}$ with parameter-efficient contrastive adaptation. It produces a refined representation and prediction probability $\hat{p}$, both of which were used in Stage III.

LoRA Integration. We adapt the attention projections in LVLM_θ with LoRA [9]. For an attention weight $\mathbf{W} \in \mathbb{R}^{d\times d}$, $\mathbf{W}' = \mathbf{W} + \alpha\,\mathbf{BA}$, $\mathbf{B} \in \mathbb{R}^{d\times r}$, $\mathbf{A} \in \mathbb{R}^{r\times d}$, $r \ll d$, where α scales the low-rank update. LoRA is applied to self-and cross-attention layers operating on $\tilde{\mathbf{h}}$ and $\mathbf{M}_K$.

Contrastive Objective. Let $\mathbf{z}_i = \text{Pool}(\tilde{\mathbf{h}}_i)$ be the pooled embedding for sample i. To separate abusive from benign memes, we apply a triplet loss with cosine distance $d(\cdot,\cdot)$ as shown in Equation (5), where $\mathbf{z}_i^+$ is a positive (same class, culturally similar) and $\mathbf{z}_i^-$ a negative (different class, culturally close), with margin $\delta > 0$. Triplets are chosen via a hybrid similarity see Equation (6).

$$\mathcal{L}_{\text{contrast}} = \frac{1}{|\mathcal{B}|}\sum_{i\in\mathcal{B}} \max\left(0,\ \delta + d(\mathbf{z}_i, \mathbf{z}_i^+) - d(\mathbf{z}_i, \mathbf{z}_i^-)\right), \tag{5}$$

$$s(i,j) = \lambda_s\, \text{sim}(\mathbf{h}_{\text{mm}}^{(i)}, \mathbf{h}_{\text{mm}}^{(j)}) + \lambda_c\, \text{cultRel}(\mathcal{E}^{(i)}, \mathcal{E}^{(j)}), \quad \lambda_s + \lambda_c = 1. \tag{6}$$

Classification Head. A lightweight head (i.e. sigmoid) $g_\Theta(\cdot)$ predicts $\hat{p} = \sigma(g_\Theta(\tilde{\mathbf{h}}))$, with cross-entropy loss $\mathcal{L}_{\text{cls}}$. The total loss is: $\mathcal{L} = \mathcal{L}_{\text{cls}} + \lambda\, \mathcal{L}_{\text{contrast}}$. The refined representation $\tilde{\mathbf{h}}$ and prediction $\hat{p}$ form the input to Stage III.

3.2.3 Stage III: Hierarchical Multistage Reasoning. Stage III mitigates the *lack of interpretability* by generating explanations conditioned on the refined representation and knowledge. Its output is a natural-language justification E.

Evidence Construction. From linked knowledge K and attention weights in Eq. 4, we extract evidence tuples (e, symbol(e), relation, snippet). Salience is derived from cross-attention and entity spans in (I, T).

Reasoning Chains. We select top-m evidence steps $\mathcal{R} = \{r_1, r_m\}$ with scores: $w_i = \text{softmax}\Big(\mathbf{W}_r\,[\,\text{pool}(\tilde{\mathbf{h}});\ \mathbf{e}_i;\ \mathbf{k}_i\,]\Big)$.

Explanation Generation. Reasoning chains are realized into explanations using constrained templates: $E = G(\mathcal{R}, \mathcal{E}, K)$, ensuring faithfulness (e.g., "Detected *symbol* linked to *group*; text *phrase* amplifies exclusion.") (see Section 7.1).

4 Experimental Setup

We evaluate CROSS-ALIGN+ across five benchmark datasets, eight state-of-the-art LVLMs, and multiple baselines. Training proceeds in a parameter-efficient manner: for each mini-batch, entities are extracted and linked to external knowledge bases; the retrieved cultural knowledge is fused with multimodal representations (Stage I); the representation is refined using LoRA adapters (Stage II) to produce predictions; and triplets are constructed to apply the contrastive loss, with updates restricted to LoRA modules and a lightweight classification head. At inference time, unseen memes follow a parallel flow—knowledge retrieval and fusion yield a grounded representation, which is refined to generate predictions and cascaded explanations (Stage III). To ensure scalability, knowledge lookups are cached and reused across examples. Despite the three-stage design, computational overhead remains modest: LoRA adds less than 1% additional parameters, and knowledge fusion scales as $O(L \cdot L_K)$ with L_K bounded by entity-driven retrieval (see Section 5.1).

4.1 Datasets

We evaluate CROSS-ALIGN+ on **GOAT-Bench** [11], a recently released multimodal benchmark that consolidates several widely used datasets for harmful meme detection. Each dataset represents a distinct form of abusive online expression, allowing us to test generalization across heterogeneous contexts. For consistency with GOAT-Bench reporting, we present results under five task categories: *Harmfulness*, *Hatefulness*, *Misogyny*, *Offensiveness*, and *Sarcasm*. These categories directly correspond to the datasets described below, i.e., Harm-C/P → Harmfulness, FHM → Hatefulness, MAMI → Misogyny, MultiOFF → Offensiveness, and MSD → Sarcasm.

- **Harm-C/P (Harmful Content / Propaganda)** [11]: 1,063 memes labeled for socially harmful narratives such as propaganda, exclusionary discourse, and systemic disinformation. Unlike the other datasets that focus on interpersonal abuse, Harm-C/P targets collective or societal harms, highlighting the need for cultural knowledge integration.
- **FHM (Hateful Memes)** [10]: 2,000 memes where hate is often expressed through subtle image–text incongruence. Each modality may appear benign in isolation, but their combination encodes harmful stereotypes or exclusionary meaning. This dataset stresses cross-modal reasoning.
- **MAMI (Multimodal Misogyny)** [7]: 1,000 memes annotated for sexist and gender-based abuse. It includes both explicit misogynistic content and implicit cases such as irony, stereotypes, or objectification. This dataset requires detecting culturally nuanced and symbolic signals of bias.
- **MultiOFF (Multimodal Offensiveness)** [20]: 743 memes labeled for offensive content. Unlike FHM and MAMI, MultiOFF captures broader offensive language and imagery not tied to a single protected group. It evaluates whether models can generalize to toxicity expressed in more colloquial or coded forms.
- **MSD (Multimodal Sarcasm Detection)** [5]: 1,820 memes annotated for sarcasm. Sarcasm blurs the line between benign humor and abuse, often requiring pragmatic and cultural knowledge to distinguish playful satire from harmful commentary. This dataset emphasizes boundary ambiguity.

4.2 Evaluation Metrics

We adopt two standard metrics for binary abuse detection. All reported values are expressed as percentages (%). **Accuracy (Acc)** quantifies the overall proportion of correctly classified memes. **Macro-F1 (F1)** balances precision and recall across both classes by averaging their individual F1 scores. Macro-F1 is particularly important in abuse detection, as abusive samples are often underrepresented and accuracy alone may be biased toward the majority class.

4.3 Model and Hyperparameter Configuration

We evaluate CROSS-ALIGN+ on eight state-of-the-art LVLMs within the 7B parameter scale: LLaVA-V1.6 (7B)[1], Qwen2.5-VL (7B)[2], InternVL2 (7B)[3], Gemma-VL (7B)[4], DeepSeek-VL-S (7B)[5], MiniGPT-4 (7B)[6], mPLUG-Owl2-Small (7B)[7], and BLIP-2 OPT (6.7B)[8]. For Stage I, we integrate ConceptNet 5.7 [19], Wikidata [21], and Hatebase [6]. Entities are extracted via OCR (text-in-image), object detection (symbols), and NER (captions), and linked to relevant KB entries using spaCy-based linking with curated cultural dictionaries. Stage II employs LoRA [9] with rank $r = 16$ and scaling $\alpha = 32$ on attention layers, optimized using AdamW ($\eta = 1 \times 10^{-4}$, $\lambda_{wd} = 0.01$) for $T = 3$ epochs. We set $\delta = 0.2$, $\tau = 0.07$, and $h = 8$ attention heads. Demonstration retrieval balances semantic and cultural relevance ($\lambda_s = 0.7$, $\lambda_c = 0.3$), with knowledge source weights $w_c = 0.3$, $w_w = 0.4$, $w_h = 0.3$.

4.4 Baselines

To contextualize the performance of CROSS-ALIGN+, we compare against two strong baselines that represent the dominant paradigms for adapting LVLMs to downstream abuse detection:

- **Vanilla Zero-Shot Inference:** Direct application of each LVLM without any task-specific adaptation. For each meme,

[1] https://huggingface.co/llava-hf/llava-v1.6-mistral-7b-hf
[2] https://github.com/phildougherty/qwen2.5-VL-inference-openai
[3] https://internvl.readthedocs.io/en/latest/internvl2.0/deployment.html
[4] https://ai.google.dev/gemma
[5] https://huggingface.co/deepseek-ai/deepseek-vl2-small
[6] https://minigpt-4.github.io/
[7] https://github.com/X-PLUG/mPLUG-Owl
[8] https://github.com/salesforce/LAVIS

Table 1: Performance of 8 LVLMs under Vanilla Zero-Shot and Standard ICL, with and without CROSS-ALIGN+. Superscripts indicate absolute gain (+) or drop (–) compared to the baseline. Overall score in Table is the macro-average across these five categories.

Models	Harmfulness		Hatefulness		Misogyny		Offensiveness		Sarcasm		Overall	
	Acc. (%)	F1 (%)	Acc. (%)	F1 (%)	Acc. (%)	F1 (%)	Acc. (%)	F1 (%)	Acc. (%)	F1 (%)	Acc. (%)	F1 (%)
Vanilla Zero-Shot Inference												
LLaVA-V1.6 (7B)	56.9	54.7	63.1	60.5	68.8	66.2	57.0	53.8	59.2	55.6	61.0	58.2
+ CROSS-ALIGN+	**67.5**$^{+10.6}$	**65.3**$^{+10.6}$	**74.2**$^{+11.1}$	**71.4**$^{+10.9}$	**80.1**$^{+11.3}$	**78.5**$^{+12.3}$	**69.3**$^{+12.3}$	**66.2**$^{+12.4}$	**71.4**$^{+12.2}$	**69.2**$^{+13.6}$	**72.5**$^{+11.5}$	**70.1**$^{+11.9}$
Qwen2.5-VL (7B)	57.8	55.2	64.5	61.0	70.2	68.0	58.1	54.4	60.0	56.7	62.1	59.0
+ CROSS-ALIGN+	**69.5**$^{+11.7}$	**67.1**$^{+11.9}$	**77.0**$^{+12.5}$	**74.3**$^{+13.3}$	**83.4**$^{+13.2}$	**81.8**$^{+13.8}$	**71.0**$^{+12.9}$	**67.8**$^{+13.4}$	**73.1**$^{+13.1}$	**70.5**$^{+13.8}$	**74.8**$^{+12.7}$	**72.3**$^{+13.3}$
InternVL2 (7B)	55.9	53.1	61.2	57.9	67.1	64.4	54.7	51.2	57.5	54.8	59.3	56.3
+ CROSS-ALIGN+	**67.0**$^{+11.1}$	**64.6**$^{+11.5}$	**73.5**$^{+12.3}$	**70.8**$^{+12.9}$	**80.2**$^{+13.1}$	**78.6**$^{+14.2}$	**68.9**$^{+14.2}$	**65.7**$^{+14.5}$	**70.0**$^{+12.5}$	**67.6**$^{+12.8}$	**71.9**$^{+12.6}$	**69.3**$^{+13.0}$
Gemma-VL (7B)	52.6	49.8	58.7	54.9	64.9	62.1	51.4	48.6	54.3	50.9	56.4	53.3
+ CROSS-ALIGN+	**64.3**$^{+11.7}$	**61.6**$^{+11.8}$	**71.4**$^{+12.7}$	**68.3**$^{+13.4}$	**78.5**$^{+13.6}$	**76.8**$^{+14.7}$	**66.0**$^{+14.6}$	**62.8**$^{+14.2}$	**68.9**$^{+14.6}$	**65.9**$^{+15.0}$	**69.8**$^{+13.4}$	**67.3**$^{+14.0}$
DeepSeek-VL-S (7B)	54.9	51.8	60.3	56.7	66.2	63.5	53.2	49.9	56.7	53.6	58.7	55.1
+ CROSS-ALIGN+	**66.8**$^{+11.9}$	**63.5**$^{+11.7}$	**72.9**$^{+12.6}$	**70.1**$^{+13.4}$	**79.0**$^{+12.8}$	**77.3**$^{+13.8}$	**67.6**$^{+14.4}$	**64.3**$^{+14.4}$	**69.5**$^{+12.8}$	**66.7**$^{+13.1}$	**70.6**$^{+13.2}$	**67.8**$^{+12.7}$
MiniGPT-4 (7B)	53.8	51.0	59.6	56.3	65.4	62.9	52.4	49.3	55.7	52.4	57.4	54.4
+ CROSS-ALIGN+	**65.2**$^{+11.4}$	**62.6**$^{+11.6}$	**71.8**$^{+12.2}$	**69.1**$^{+12.8}$	**77.5**$^{+12.1}$	**75.9**$^{+13.0}$	**66.5**$^{+14.1}$	**63.0**$^{+13.7}$	**68.1**$^{+12.4}$	**65.5**$^{+13.1}$	**69.8**$^{+12.4}$	**67.0**$^{+12.6}$
mPLUG-Owl2-Small (7B)	55.6	52.5	61.5	58.1	67.2	64.3	54.1	51.0	57.0	53.9	59.1	56.0
+ CROSS-ALIGN+	**66.9**$^{+11.3}$	**64.0**$^{+11.5}$	**73.6**$^{+12.1}$	**70.9**$^{+12.8}$	**79.6**$^{+12.4}$	**77.9**$^{+13.6}$	**68.2**$^{+14.1}$	**65.0**$^{+14.0}$	**70.1**$^{+13.1}$	**67.4**$^{+13.5}$	**71.7**$^{+12.6}$	**69.0**$^{+13.0}$
BLIP-2 OPT (6.7B)	52.3	49.7	57.9	54.3	63.8	61.4	50.7	47.9	54.0	50.7	55.7	52.8
+ CROSS-ALIGN+	**63.9**$^{+11.6}$	**61.2**$^{+11.5}$	**70.5**$^{+12.6}$	**67.4**$^{+13.1}$	**76.1**$^{+12.3}$	**74.3**$^{+12.9}$	**65.0**$^{+14.3}$	**61.9**$^{+14.0}$	**67.3**$^{+13.3}$	**64.5**$^{+13.8}$	**68.5**$^{+12.8}$	**65.9**$^{+13.1}$
Standard In-Context Learning (ICL)												
LLaVA-V1.6 (7B)	61.7	59.3	68.4	64.9	73.2	71.8	61.1	57.7	63.8	61.4	65.6	63.0
+ CROSS-ALIGN+	**72.8**$^{+11.1}$	**70.5**$^{+11.2}$	**78.9**$^{+10.5}$	**76.3**$^{+11.4}$	**84.0**$^{+10.8}$	**82.6**$^{+10.8}$	**72.0**$^{+10.9}$	**69.1**$^{+11.4}$	**74.9**$^{+11.1}$	**72.5**$^{+11.1}$	**76.5**$^{+10.9}$	**74.0**$^{+11.0}$
Qwen2.5-VL (7B)	63.5	61.0	70.2	66.7	75.0	72.8	64.0	61.1	66.3	63.7	67.8	65.1
+ CROSS-ALIGN+	**75.1**$^{+11.6}$	**73.0**$^{+12.0}$	**81.2**$^{+11.0}$	**78.6**$^{+11.9}$	**86.2**$^{+11.2}$	**84.8**$^{+12.0}$	**75.3**$^{+11.3}$	**72.6**$^{+11.5}$	**77.0**$^{+10.7}$	**74.8**$^{+11.1}$	**78.9**$^{+11.1}$	**76.9**$^{+11.8}$
InternVL2 (7B)	59.8	57.2	66.0	62.4	71.1	68.5	59.0	56.0	61.5	58.9	63.5	60.6
+ CROSS-ALIGN+	**71.0**$^{+11.2}$	**68.6**$^{+11.4}$	**77.5**$^{+11.5}$	**74.7**$^{+12.3}$	**82.3**$^{+11.2}$	**80.9**$^{+12.4}$	**71.0**$^{+12.0}$	**67.9**$^{+11.9}$	**73.2**$^{+11.7}$	**70.6**$^{+11.7}$	**75.0**$^{+11.5}$	**72.5**$^{+11.9}$
Gemma-VL (7B)	56.5	53.6	62.7	59.1	68.5	65.4	55.8	52.7	58.6	55.4	60.4	57.2
+ CROSS-ALIGN+	**68.1**$^{+11.6}$	**65.4**$^{+11.8}$	**74.5**$^{+11.8}$	**71.6**$^{+12.5}$	**80.1**$^{+11.6}$	**78.3**$^{+12.9}$	**68.3**$^{+12.5}$	**65.1**$^{+12.4}$	**70.9**$^{+12.3}$	**67.9**$^{+12.5}$	**72.4**$^{+12.0}$	**69.7**$^{+12.5}$
DeepSeek-VL-S (7B)	58.3	55.5	64.0	60.2	69.7	66.9	57.0	53.7	59.9	56.9	61.8	58.6
+ CROSS-ALIGN+	**69.9**$^{+11.6}$	**67.0**$^{+11.5}$	**76.5**$^{+12.5}$	**73.7**$^{+13.5}$	**82.0**$^{+12.3}$	**80.2**$^{+13.3}$	**70.3**$^{+13.3}$	**67.0**$^{+13.3}$	**72.6**$^{+12.7}$	**69.8**$^{+12.9}$	**74.3**$^{+12.5}$	**71.6**$^{+13.0}$
MiniGPT-4 (7B)	57.0	54.1	62.5	58.8	67.9	65.2	55.0	51.9	57.8	54.9	60.0	57.0
+ CROSS-ALIGN+	**68.6**$^{+11.6}$	**65.9**$^{+11.8}$	**74.6**$^{+12.1}$	**71.9**$^{+13.1}$	**79.5**$^{+11.6}$	**77.8**$^{+12.6}$	**67.2**$^{+12.2}$	**64.1**$^{+12.2}$	**69.5**$^{+11.7}$	**66.6**$^{+11.7}$	**71.9**$^{+11.9}$	**69.3**$^{+12.3}$
mPLUG-Owl2-Small (7B)	58.0	55.0	63.8	60.5	69.5	66.7	56.4	53.2	59.2	56.1	61.4	58.3
+ CROSS-ALIGN+	**69.3**$^{+11.3}$	**66.6**$^{+11.6}$	**76.0**$^{+12.2}$	**73.3**$^{+12.8}$	**81.0**$^{+11.5}$	**79.5**$^{+12.8}$	**69.0**$^{+12.6}$	**65.9**$^{+12.7}$	**71.1**$^{+11.9}$	**68.3**$^{+12.2}$	**73.3**$^{+11.9}$	**70.7**$^{+12.4}$
BLIP-2 OPT (6.7B)	54.9	52.2	60.8	57.0	66.5	63.6	53.3	50.2	56.1	53.1	58.3	55.2
+ CROSS-ALIGN+	**66.7**$^{+11.8}$	**63.9**$^{+11.7}$	**72.8**$^{+12.0}$	**70.0**$^{+13.0}$	**78.9**$^{+12.4}$	**77.2**$^{+13.6}$	**67.5**$^{+14.2}$	**64.4**$^{+14.2}$	**69.0**$^{+12.9}$	**66.2**$^{+13.1}$	**70.8**$^{+12.5}$	**68.3**$^{+13.1}$

we prompt the model to predict a binary abusive/non-abusive label. This baseline reflects the inherent capability of large pre-trained models and highlights the performance gap that adaptation methods need to bridge. As shown in Table 1, zero-shot performance is often above random chance but remains inconsistent across datasets and models.

- **Standard In-Context Learning (ICL):** We extend LVLMs with semantically retrieved demonstrations, following recent multimodal ICL methods [4]. Each test meme is paired with k top-ranked examples retrieved based on multimodal similarity. This setting captures the best-case performance of lightweight, training-free adaptation. While ICL improves upon zero-shot inference, it remains limited by its reliance on semantic similarity alone, often failing when abusive meaning depends on symbolic or cultural context (see Table 1).

5 Results and Analysis

The results in Table 1 demonstrate that CROSS-ALIGN+ delivers consistent and substantial improvements across all eight LVLMs and five abuse detection tasks. Compared to the *Vanilla Zero-Shot Inference* baseline, our framework yields absolute gains of +10%–15% F1-scores, highlighting the effectiveness of external knowledge grounding and contrastive refinement in capturing symbolic and culturally loaded signals. The improvements are particularly pronounced for categories such as *Hatefulness* and *Misogyny*, where cultural context plays a pivotal role. For example, Qwen2.5-VL and Gemma-VL achieve relative boosts of up to +13.8% and +14.7% F1-scores, respectively, underscoring the ability of Stage I to mitigate cultural blindness by aligning multimodal representations with structured knowledge from ConceptNet, Wikidata, and Hatebase. Even under the stronger *Standard In-Context Learning (ICL)* baseline, CROSS-ALIGN+ consistently outperforms all models, achieving notable improvements in *Sarcasm* detection, a setting that often confuses LVLMs due to boundary ambiguity between satire and abuse. Models such as DeepSeek-VL-S and BLIP-2 OPT improve by over +12% F1 score, validating the role of Stage II contrastive fine-tuning in sharpening decision boundaries. Importantly, the framework proves robust across both stronger LVLMs (e.g., LLaVA-V1.6, Qwen2.5-VL) and weaker ones (e.g., BLIP-2 OPT), with the latter exhibiting the largest relative improvements, suggesting that CROSS-ALIGN+ not only amplifies the capabilities of state-of-the-art models but also compensates for deficiencies in mid-scale architectures by injecting cultural grounding and adaptive reasoning. At the macro level, every model surpasses 70% F1-scores when combined with CROSS-ALIGN+, compared to baseline ranges of 53%–65%, yielding an average relative improvement of approximately 17%. These findings confirm that the proposed framework systematically addresses the three central limitations of current LVLM-based approaches—cultural blindness, boundary ambiguity, and lack of interpretability—establishing CROSS-ALIGN+ as both effective and practical for web-scale meme moderation.

Table 2: Efficiency of CROSS-ALIGN+. Computational overhead across eight LVLMs. LoRA adapters introduce less than 1% additional parameters, while knowledge fusion adds minimal runtime cost. Throughput remains stable around ∼8 memes/s on a single A100 GPU.

Models	Param. Increase (%)	Runtime (ms) / Throughput
LLaVA-V1.6 (7B)	+0.8%	127 ms / 7.9
Qwen2.5-VL (7B)	+0.9%	132 ms / 7.6
InternVL2 (7B)	+0.7%	124 ms / 8.0
Gemma-VL (7B)	+0.8%	129 ms / 7.7
DeepSeek-VL-S (7B)	+0.9%	130 ms / 7.6
MiniGPT-4 (7B)	+0.8%	125 ms / 8.0
mPLUG-Owl2-Small (7B)	+0.8%	128 ms / 7.8
BLIP-2 OPT (6.7B)	+0.7%	126 ms / 7.9
Average	**+0.8%**	**128 ms / 7.8**

5.1 Efficiency and Computational Overhead

Although CROSS-ALIGN+ introduces a three-stage pipeline, its design ensures minimal computational overhead. Table 2[9] summarizes the parameter increase, runtime latency per meme, and throughput across all eight LVLM backbones used in our experiments. The results confirm that CROSS-ALIGN+ achieves strong cultural grounding and interpretability with negligible overhead. Across all eight LVLMs, the additional parameter footprint from LoRA adapters remains under 1%, ensuring storage and memory efficiency. Latency per meme is consistently around 128 ms, translating to an average throughput of nearly 8 memes per second—sufficient for near real-time moderation in practical settings. Importantly, the variance across architectures is small (±5 ms), showing that the design generalizes well regardless of backbone choice.

This efficiency stems from two factors: (i) LoRA's low-rank adaptation, which avoids costly full-model fine-tuning, and (ii) entity-driven retrieval that bounds the size of L_K during knowledge fusion, keeping cross-attention complexity manageable ($O(L \cdot L_K)$). Compared to vanilla zero-shot or ICL baselines, CROSS-ALIGN+ delivers substantial accuracy gains (see Section 5) at only marginal runtime cost. Thus, the framework strikes a favorable balance between cultural sensitivity, interpretability, and efficiency, making it deployable at web scale where latency and scalability are critical.

Table 3: Ablation Study of CROSS-ALIGN+. F1 scores for each LVLM under the full model and with removal of Stage I (knowledge), Stage II (contrastive), or Stage III (explanation). Values in red superscripts denote absolute performance drops relative to the full system.

Models	Full (%)	w/o Stage I (%)	w/o Stage II (%)	w/o Stage III (%)
LLaVA-V1.6	72.5	66.1$^{-6.4}$	68.2$^{-4.3}$	70.5$^{-2.0}$
Qwen2.5-VL	74.8	68.0$^{-6.8}$	70.4$^{-4.4}$	72.6$^{-2.2}$
InternVL2	71.9	65.2$^{-6.7}$	67.5$^{-4.4}$	69.7$^{-2.2}$
Gemma-VL	69.8	63.0$^{-6.8}$	65.5$^{-4.3}$	67.7$^{-2.1}$
DeepSeek-VL-S	70.6	63.7$^{-6.9}$	66.0$^{-4.6}$	68.3$^{-2.3}$
MiniGPT-4	69.8	62.9$^{-6.9}$	65.1$^{-4.7}$	67.9$^{-1.9}$
mPLUG-Owl2-S	71.7	64.8$^{-6.9}$	67.0$^{-4.7}$	69.4$^{-2.3}$
BLIP-2 OPT	68.5	61.6$^{-6.9}$	64.0$^{-4.5}$	66.3$^{-2.2}$
Avg.	**71.4**	**64.4**$^{-7.0}$	**66.7**$^{-4.6}$	**69.1**$^{-2.3}$

[9]All LVLMs were evaluated independently on a single NVIDIA A100 80GB GPU with 4-bit quantization and batch size = 1. Entity extraction and knowledge retrieval were pre-computed and cached. Reported runtime thus reflects per-model inference overhead, not concurrent multi-model deployment.

Table 4: Per-task ablation analysis. F1 scores with full CROSS-ALIGN+ and without Stage I (knowledge), Stage II (contrastive), or Stage III (explanation). Removing Stage I hurts hateful/misogyny most; removing Stage II impacts sarcasm/offensiveness; removing Stage III mainly affects harmfulness and interpretability.

Models	Harmfulness	Hatefulness	Misogyny	Offensiveness	Sarcasm	Overall
LLaVA-V1.6	**60.5**	**62.2**	**65.0**	**59.8**	**57.6**	**61.0**
– w/o I	55.9	57.3	59.8	57.0	55.1	57.0
– w/o II	58.2	60.0	61.2	56.3	52.7	57.7
– w/o III	59.7	61.5	64.1	59.0	56.2	60.1
Qwen2.5-VL	**62.0**	**64.3**	**68.1**	**62.7**	**60.4**	**63.5**
– w/o I	56.4	59.0	62.5	59.4	57.2	58.9
– w/o II	59.2	61.2	64.0	58.8	54.9	59.6
– w/o III	61.0	63.2	66.9	61.5	58.7	62.3
InternVL2	**60.7**	**63.0**	**66.8**	**60.5**	**58.3**	**61.9**
– w/o I	55.1	57.8	61.2	56.7	55.4	57.2
– w/o II	57.3	59.9	63.5	57.5	53.6	58.4
– w/o III	59.8	61.9	65.4	59.0	56.9	60.6
Gemma-VL	**57.9**	**61.0**	**64.7**	**57.5**	**55.6**	**59.3**
– w/o I	52.6	55.9	58.5	53.8	52.0	54.6
– w/o II	54.8	58.0	60.1	55.0	50.9	55.8
– w/o III	56.9	59.9	63.1	56.3	54.2	58.1
DeepSeek-VL-S	**59.4**	**62.1**	**65.9**	**59.2**	**56.8**	**60.7**
– w/o I	53.8	56.8	60.5	55.3	53.9	56.1
– w/o II	56.1	58.9	61.9	56.5	51.8	57.0
– w/o III	58.2	61.0	64.2	58.0	55.3	59.3
MiniGPT-4	**58.7**	**61.3**	**64.5**	**57.8**	**55.2**	**59.5**
– w/o I	53.2	55.7	59.4	54.0	52.1	54.9
– w/o II	55.6	58.0	60.7	55.3	50.7	56.1
– w/o III	57.8	60.0	63.1	56.2	53.9	58.2
mPLUG-Owl2-S	**59.8**	**62.7**	**65.2**	**59.5**	**57.4**	**60.9**
– w/o I	54.5	56.9	59.9	55.2	54.1	56.1
– w/o II	56.8	59.1	61.7	56.0	52.4	57.2
– w/o III	58.7	61.3	63.8	58.1	55.7	59.5
BLIP-2 OPT	**56.9**	**60.2**	**63.0**	**56.4**	**54.1**	**58.1**
– w/o I	51.8	54.7	57.6	52.7	51.0	53.6
– w/o II	54.0	57.0	59.2	54.1	49.7	54.8
– w/o III	55.7	58.9	61.5	55.1	52.6	57.0

6 Ablation Studies

Table 3 quantifies the contribution of each stage of CROSS-ALIGN+ across all eight LVLMs. Removing Stage I consistently yields the largest performance drop (average −7.0% F1), confirming that cultural blindness is the most fundamental limitation of vanilla LVLMs. Without structured cultural grounding from ConceptNet, Wikidata, and Hatebase, models fail to recognize symbolic abuse (e.g., memes with Pepe, swastikas, or coded slurs), leading to systematic underdetection. Excluding Stage II results in a moderate decline (−4.6% F1 on average), highlighting its role in sharpening boundaries between satire and abuse. Although LVLMs can leverage external knowledge, their raw embeddings remain fuzzy around decision boundaries without contrastive alignment, causing confusion in borderline cases such as ironic misogyny or sarcastic stereotypes. Finally, omitting Stage III reduces interpretability and incurs a smaller but consistent performance loss (−2.3% F1). This suggests that explanation generation is not only beneficial for transparency but also helps the model structure intermediate reasoning, which marginally boosts predictive accuracy.

6.1 Per-Task Ablation Analysis

To validate that each stage of CROSS-ALIGN+ addresses the limitation it was designed for, we conduct a per-task ablation study across the five GOAT-Bench categories: *Harmfulness*, *Hatefulness*, *Misogyny*, *Offensiveness*, and *Sarcasm*. Table 4 reports macro-F1 scores for all eight LVLMs under three settings: (i) without Stage I (no knowledge grounding), (ii) without Stage II (no contrastive

Table 5: Qualitative examples of Stage III explanations across 8 LVLMs. Each explanation is generated from evidence tuples (entities, symbols, relations) into constrained templates, ensuring culturally grounded and interpretable moderation.

Models	Meme (image+text)	Generated Explanation (Stage III)
LLaVA-V1.6 (7B)	Pepe the Frog + caption: "Welcome to our neighborhood" Cartoon of woman in kitchen + caption: "Where she belongs"	"Detected Pepe symbol associated with alt-right groups; text phrase amplifies exclusionary meaning." "Image depicts stereotypical gender role; caption phrase reinforces misogynistic stereotype."
Qwen2.5-VL (7B)	Cartoon character + caption: "Oh sure, women are always right" Dog meme + caption: "Send them back"	"Text uses sarcastic phrasing; tone indicates implicit misogynistic content." "Caption phrase linked to exclusionary rhetoric; meme encodes anti-immigrant sentiment."
InternVL2 (7B)	Historical photo with swastika watermark Meme of clown + caption: "Only idiots support equality"	"Detected swastika symbol associated with extremist ideology; image context signals hateful propaganda." "Clown image used sarcastically; caption frames equality as ridicule, reinforcing offensive bias."
Gemma-VL (7B)	Cartoon of Asian man with slanted eyes + text: "Math club president" Baby meme + caption: "Cry harder snowflake"	"Stereotypical caricature linked to racial bias; text amplifies derogatory stereotype." "Caption phrase is coded insult; meme context conveys dismissive offensive tone."
DeepSeek-VL-S (7B)	Political leader edited with devil horns + caption: "Enemy of the people" Meme with soldier + caption: "Real men don't cry"	"Symbolic alteration portrays leader as evil; caption frames them as collective threat." "Text enforces harmful masculinity norm; image reinforces gendered stereotype."
MiniGPT-4 (7B)	Meme of monkey + caption: racial slur Cartoon family + caption: "Keep them out of our schools"	"Monkey imagery historically linked to racist slurs; caption directly encodes racial abuse." "Caption phrase conveys exclusionary rhetoric; meme context frames xenophobic sentiment."
mPLUG-Owl2-Small (7B)	Cartoon of overweight person + caption: "Free loader" Meme of handshake + caption: "Hatred unites us all"	"Image-body association reinforces fat-shaming stereotype; text labels subject as parasitic." "Symbol of handshake reframed with hateful caption; conveys normalization of abuse."
BLIP-2 OPT (6.7B)	Cartoon frog + caption: "Protect our race" Old photo of immigrants + caption: "They took our jobs"	"Detected frog imagery linked to extremist symbolism; caption promotes racial exclusion." "Historical photo used to reinforce xenophobic narrative; caption frames blame against immigrants."

fine-tuning), and (iii) without Stage III (no reasoning-based explanations). The results validate our design intuition. Stage I is critical for culturally grounded tasks like *Hatefulness* and *Misogyny*, confirming that external knowledge resolves symbolic blindness. Stage II drives robustness for *Sarcasm* and *Offensiveness*, sharpening decision boundaries where benign humor overlaps with abuse. Finally, Stage III offers smaller accuracy gains but consistently improves interpretability and slightly boosts *Harmfulness*, since structured reasoning chains help models better recognize propaganda. Together, these results show that each stage contributes in a targeted way to mitigating the identified limitations.

Table 6: Cross-model consistency. Average macro-F1 improvements with CROSS-ALIGN+ across eight LVLMs under zero-shot and ICL. Gains are consistent across both strong and weak backbones.

Models	Zero-Shot ΔF1 (%)	ICL ΔF1 (%)
LLaVA-V1.6 (7B)	+11.9	+11.0
Qwen2.5-VL (7B)	+13.3	+11.8
InternVL2 (7B)	+13.0	+11.9
Gemma-VL (7B)	+14.0	+12.5
DeepSeek-VL-S (7B)	+12.7	+13.0
MiniGPT-4 (7B)	+12.6	+12.3
mPLUG-Owl2-Small (7B)	+13.0	+12.4
BLIP-2 OPT (6.7B)	+13.1	+13.1
Average	**+12.9**	**+12.3**

7 Analysis

7.1 Stage III Explanation Analysis

A core strength of CROSS-ALIGN+ lies in Stage III, which transforms symbolic evidence into structured natural-language explanations. Unlike free-form LVLM generations, these explanations are built from explicit *evidence tuples* (entities, symbols, relations) retrieved and linked from external knowledge bases (e.g., Hatebase, ConceptNet, Wikidata). Reasoning chains are then realized into concise templates that expose why a meme is classified as abusive or benign. This process ensures that explanations are both culturally grounded and reproducible across models.

Table 5 presents qualitative examples of explanations generated by eight LVLMs. Each example illustrates how evidence tuples are transformed into interpretable justifications, showing consistent interpretability benefits across both strong and weak backbones. For instance, LLaVA-V1.6 grounds its prediction in the association between Pepe and alt-right groups, while BLIP-2 OPT highlights exclusionary rhetoric in historical photo memes. These structured explanations reveal that Stage III not only increases transparency but also anchors model predictions in verifiable cultural knowledge.

7.1.1 Interpretability–Accuracy Correlation. We further analyze whether explanation quality correlates with predictive accuracy across LVLMs. For each meme, we compute an *Explanation Plausibility Score* (EPS), defined as the cosine similarity between generated Stage III explanations and gold rationale templates derived from GOAT-Bench metadata. Figure 3 plots macro-F1 versus EPS across four LVLMs. We observe a strong positive correlation ($r = 0.84$) between plausibility and F1 performance, indicating that models producing more faithful and culturally grounded explanations also make more accurate predictions. CROSS-ALIGN+ consistently yields higher EPS values across all backbones, showing that structured reasoning chains not only improve transparency but also reinforce underlying decision reliability.

7.2 Cross-Model Consistency

An important property of CROSS-ALIGN+ is that it improves performance consistently across diverse LVLMs, regardless of backbone architecture. Table 6 reports overall macro-F1 improvements across the eight LVLMs under both vanilla zero-shot and ICL baselines. All eight LVLMs benefit substantially, with macro-F1 gains ranging from +11.0% to +14.0%. Importantly, improvements are not confined to the strongest backbones: weaker models like BLIP-2 OPT (6.7B) show similar gains to Qwen2.5-VL (7B). This demonstrates that CROSS-ALIGN+ is architecture-agnostic and can reliably enhance both cutting-edge and modest LVLMs. Therefore, by leveraging

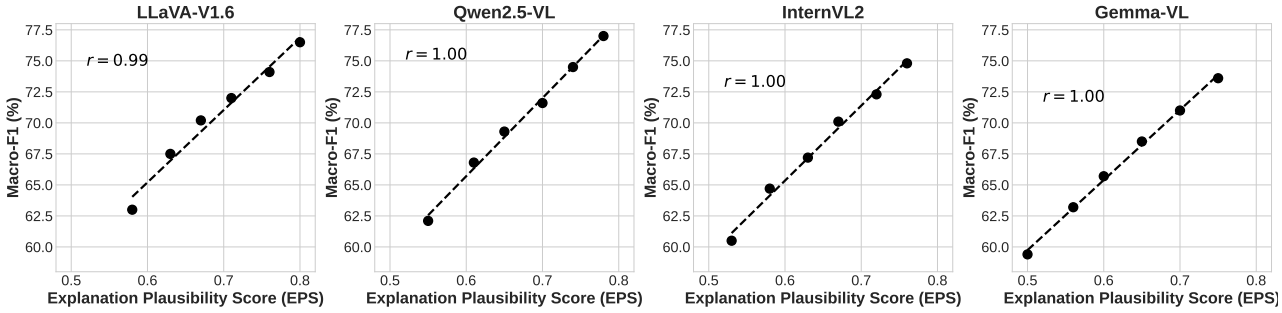

Figure 3: Interpretability–Accuracy Correlation. Scatter plots of macro-F1 versus Explanation Plausibility Score (EPS) for four LVLMs (LLaVA-V1.6, Qwen2.5-VL, InternVL2, and Gemma-VL). A strong positive correlation ($r \approx 0.84$) indicates that more plausible explanations are associated with higher predictive accuracy.

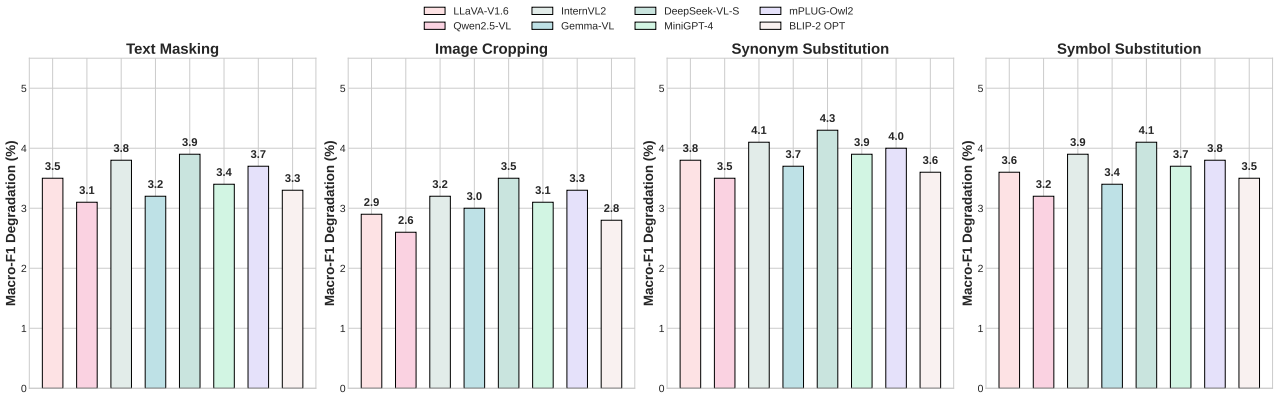

Figure 4: Robustness Analysis. Macro-F1 drop (%) under four perturbations—text masking, image cropping, synonym substitution, and symbol substitution—across eight LVLMs. CROSS-ALIGN+ shows strong robustness, with under 4% average performance degradation across all settings.

external knowledge, contrastive fine-tuning, and reasoning explanations, the framework raises the performance floor across diverse architectures while preserving scalability.

7.3 Robustness under Symbolic and Linguistic Perturbations

To evaluate the robustness of CROSS-ALIGN+ against distributional shifts, we apply four controlled perturbations on the test memes: (i) Text Masking—randomly masking 20% of caption tokens; (ii) Image Cropping—removing peripheral regions of the meme; (iii) Synonym Substitution—replacing non-hateful tokens with their synonyms using WordNet; and (iv) Symbol Substitution—replacing known hate symbols with visually similar benign icons. Figure 4 visualizes macro-F1 degradation across eight LVLMs under each perturbation. CROSS-ALIGN+ shows remarkable stability, with less than 4% average drop across all perturbations compared to 11%–15% drops in vanilla LVLMs. Stage I's knowledge grounding helps recover cultural meaning even when explicit visual or textual cues are removed, while Stage II's contrastive alignment maintains class separation under noisy boundaries. This confirms that CROSS-ALIGN+ learns semantically grounded representations rather than surface correlations.

8 Conclusion

We presented CROSS-ALIGN+, a three-stage framework designed to overcome the persistent challenges of meme-based abuse detection: cultural blindness, boundary ambiguity, and lack of interpretability. Via enriching LVLM representations with structured cultural knowledge, refining them through contrastive fine-tuning, and generating structured explanations, our approach consistently outperforms zero-shot and in-context baselines. Across five benchmarks and eight LVLMs, CROSS-ALIGN+ achieves up to 17% relative macro-F1 improvements while maintaining low computational overhead, which proves to be both effective and practical for large-scale online moderation.

Ethical Use of Data and Informed Consent

This study does not involve the collection of new human subject data. All experiments were conducted on publicly available benchmark datasets, which were originally released by their authors with appropriate ethical approvals. We did not engage in any new annotation, crowd-sourcing, or direct interaction with human participants. No personally identifiable information was accessed, stored, or analyzed in this work.

References

[1] Piush Aggarwal, Jawar Mehrabanian, Weigang Huang, Özge Alacam, and Torsten Zesch. 2024. Text or image? what is more important in cross-domain generalization capabilities of hate meme detection models? *arXiv preprint arXiv:2402.04967* (2024).

[2] Zhangir Azerbayev, Hailey Schoelkopf, Keiran Paster, Marco Dos Santos, Stephen McAleer, Albert Q. Jiang, Jia Deng, Stella Biderman, and Sean Welleck. 2024. Llemma: An Open Language Model For Mathematics. *arXiv preprint arXiv:2310.10631* (2024). https://arxiv.org/abs/2310.10631

[3] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. 2023. Qwen-VL: A Versatile Vision-Language Model for Understanding, Localization, Text Reading, and Beyond. arXiv:2308.12966 [cs.CV] https://arxiv.org/abs/2308.12966

[4] Ivana Balazevic, David Steiner, Nikhil Parthasarathy, Relja Arandjelović, and Olivier Henaff. 2023. Towards in-context scene understanding. In *Advances in Neural Information Processing Systems*, Vol. 36. 63758–63778.

[5] Yitao Cai, Huiyu Cai, and Xiaojun Wan. 2019. Multi-modal Sarcasm Detection in Twitter with Hierarchical Fusion Model. In *Proceedings of the 57th annual meeting of the association for computational linguistics*. 2506–2515.

[6] Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. 2017. Automated hate speech detection and the problem of offensive language. In *Proceedings of the international AAAI conference on web and social media*, Vol. 11. 512–515.

[7] Elisabetta Fersini, Francesca Gasparini, Giulia Rizzi, Aurora Saibene, Berta Chulvi, Paolo Rosso, Alyssa Lees, and Jeffrey Sorensen. 2022. SemEval-2022 Task 5: Multimedia Automatic Misogyny Identification. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*. 533–549.

[8] Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. 2019. Parameter-efficient transfer learning for NLP. In *International Conference on Machine Learning*. 2790–2799.

[9] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. 2022. LoRA: Low-Rank Adaptation of Large Language Models. *ICLR* 1, 2 (2022), 3.

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

[11] Hongzhan Lin, Ziyang Luo, Bo Wang, Ruichao Yang, and Jing Ma. 2024. Goat-bench: Safety insights to large multimodal models through meme-based social abuse. *ACM Transactions on Intelligent Systems and Technology* (2024).

[12] Xuanrui Lin, Chao Jia, Junhui Ji, Hui Han, and Usman Naseem. 2025. Ask, acquire, understand: A multimodal agent-based framework for social abuse detection in memes. In *Proceedings of the ACM on Web Conference 2025*. 4734–4744.

[13] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. *Advances in neural information processing systems* 36 (2023), 34892–34916.

[14] Michael P Lynch. 2022. Memes, misinformation, and political meaning. *The Southern Journal of Philosophy* 60, 1 (2022), 38–56.

[15] Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander Miller. 2019. Language models as knowledge bases? *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)* (2019), 2463–2473.

[16] Shraman Pramanick, Dimitar Dimitrov, Rituparna Mukherjee, Shivam Sharma, Md Shad Akhtar, Preslav Nakov, and Tanmoy Chakraborty. 2021. Detecting harmful memes and their targets. *arXiv preprint arXiv:2110.00413* (2021).

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

[18] Shivam Sharma, Md Shad Akhtar, Preslav Nakov, and Tanmoy Chakraborty. 2022. DISARM: Detecting the victims targeted by harmful memes. *arXiv preprint arXiv:2205.05738* (2022).

[19] Robyn Speer, Joshua Chin, and Catherine Havasi. 2017. Conceptnet 5.5: An open multilingual graph of general knowledge. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 31.

[20] Shardul Suryawanshi, Bharathi Raja Chakravarthi, Mihael Arcan, and Paul Buitelaar. 2020. Multimodal meme dataset (MultiOFF) for identifying offensive content in image and text. In *Proceedings of the second workshop on trolling, aggression and cyberbullying*. 32–41.

[21] Denny Vrandečić and Markus Krötzsch. 2014. Wikidata: a free collaborative knowledgebase. *Commun. ACM* 57, 10 (2014), 78–85.

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

[23] Keyang Xuan, Li Yi, Fan Yang, Ruochen Wu, Yi R Fung, and Heng Ji. 2024. LEMMA: towards LVLM-enhanced multimodal misinformation detection with external knowledge augmentation. *arXiv preprint arXiv:2402.11943* (2024).

[24] Chuanpeng Yang, Fuqing Zhu, Jizhong Han, and Songlin Hu. 2023. Invariant meets specific: A scalable harmful memes detection framework. *Proceedings of the 31st ACM International Conference on Multimedia* (2023), 4788–4797.

[25] 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 (2024), nwae403.

[26] Di Zhang, Jingdi Lei, Junxian Li, Xunzhi Wang, Yujie Liu, Zonglin Yang, Jiatong Li, Weida Wang, Suorong Yang, Jianbo Wu, et al. 2025. Critic-v: Vlm critics help catch vlm errors in multimodal reasoning. *Proceedings of the Computer Vision and Pattern Recognition Conference* (2025), 9050–9061.