

Small Cues, Big Consequences: Learning Pivotal Cues for Multimodal Meme Classification

Akshit Sharma and Prashant W. Patil

CVPR Lab, MFSDSAI

Indian Institute of Technology Guwahati

akshitsharma.rs@gmail.com pwpatil@iitg.ac.in

Abstract

Memes often derive their harmful, hateful, or sarcastic meaning from small but decisive visual, textual, or cross-modal cues. Existing multimodal classifiers can miss such evidence when relying mainly on global image-text representations. We introduce MemeCF, a cue-focused benchmark of 9,895 memes across harm, hate, and sarcasm, with annotations identifying the modality and rationale of the pivotal evidence. We also propose MemePIVOT, a local-global architecture for meme classification. MemePIVOT uses frozen CLIP features, unbalanced optimal transport to align words with image patches while allowing irrelevant evidence to remain unmatched, and an evidential fusion head to combine local grounding with global meme context under uncertainty. Experiments on HarMeme, PrideMM, and MemeCF show consistent gains over strong text-only, image-only, multimodal, and vision-language baselines. Cross-dataset and ablation results further show that explicit pivotal-evidence modeling improves robustness and contributes meaningfully beyond global multimodal representations. Our code and dataset are publicly available at <https://github.com/AkshitSharma1/MemePIVOT>

1 Introduction

An internet meme is rarely the sum of an image and a caption. Its meaning often emerges from the tension between visual content, overlaid text, cultural references, symbols, humor, and social context. Like a punchline in comedy, a small visual or textual element can invert the meaning of an entire meme. A single word may redirect the target of a joke, a background symbol may introduce hateful meaning, a gesture may signal harm, or a subtle mismatch between image and text may turn an otherwise neutral post into sarcasm. This makes memes compact but highly expressive multimodal artifacts whose interpretation depends not only on what is present, but on which details matter.

The growing influence of memes has made their automatic understanding an important problem for online safety and content moderation. Memes shape political discourse, reinforce social identities, mobilize communities, and influence public opinion. At the same time, they are increasingly used to spread hateful narratives, harmful stereotypes, targeted harassment, sarcasm-driven abuse, and other forms of online harm. Detecting such content is challenging because harmful meaning is often indirect, culturally grounded, and visually sparse. A meme may appear benign when viewed globally, while a small phrase, object, symbol, facial expression, or cross-modal relation determines whether it is hateful, harmful, sarcastic, or benign.

This sparse-evidence structure creates a central challenge for existing multimodal models. Many vision-language systems rely primarily on global image-text representations, which are effective for capturing broad semantic alignment but can overlook the fine-grained evidence that drives meme interpretation. In many cases, most of the image and text provide context, while only a small set of pivotal cues determines the final label. As a result, a model may predict the correct class for the wrong reason, rely on superficial correlations, or fail when the decisive cue is visually small, semantically subtle, or surrounded by distracting context. Reliable meme classification therefore requires more than recognizing the overall scene or reading the full caption; it requires identifying the pivotal visual, textual, and cross-modal evidence that supports the decision.

To study this problem, we introduce MEMECF, a cue-focused benchmark for fine-grained meme classification. MEMECF contains 9,895 memes annotated across three socially important subtasks: hate, harm, and sarcasm. These subtasks capture three semantically distinct forms of socially consequential interpretation: identity-directed hostility, threat or injury, and pragmatic incongruity. Unlike

Work	Source / Context	Multi-task	Multi-domain	Cue / CF	Size	Context / Tasks
DisinfoMeme (Qu et al., 2022)	Reddit; COVID-19, BLM, Veganism	✗	✓	✗	1,170	Disinformation memes: COVID-19, BLM, Veganism
(Tanaka et al., 2022)	Meme websites; general	✗	NR	✗	7,500	Humor in general memes
HMC (Kiela et al., 2020)	Self-generated memes	✗	NR	○*	10,000	Hateful meme detection
MultiOFF (Suryawanshi et al., 2020)	Social media; U.S. election	✗	✗	✗	743	Offensive meme detection; U.S. election
Harm-C (Pramanick et al., 2021a)	Google Images; COVID-19	○	✗	✗	3,544	Harmful COVID-19 memes and targets
Harm-P (Pramanick et al., 2021b)	Google Images; U.S. politics	○	✗	✗	3,552	Harmful U.S. politics memes and targets
CrisisHateMM (Bhandari et al., 2023)	Twitter, FB, Reddit; war	○	✗	✗	4,723	Russia–Ukraine war hate/target classification
PrideMM (Shah et al., 2024)	FB, Twitter, Reddit; LGBTQ+	✓	✗	✗	5,063	LGBTQ+ movement: hate, target, stance, humor
MultiBully (Maity et al., 2022)	Twitter, Reddit; code-mixed	✓	✗	✗	5,854	Code-mixed cyberbullying: bully, sentiment, emotion, sarcasm, harmfulness
MMSD2.0 (Qin et al., 2023)	Twitter; sarcasm memes	✗	NR	○**	~24.3K	Multimodal sarcasm detection
GuardHarMem (El-amrany et al., 2025)	Multiplatform harmful memes	○	○	✗	~16.6K	–
GOAT-Bench (Lin et al., 2026)	Curated meme benchmark	✓	✓	✗	6K+	Hatefulness, misogyny, offensiveness, sarcasm, harmfulness
MemeCF (Ours)	Curated multi-source benchmark	✓	✓	✓	9,895	Counterfactual meme understanding across hate, harm, and sarcasm

Table 1: Comparison of MemeCF with existing meme benchmarks. CF = counterfactual-style supervision; ✓ = yes; ✗ = no; ○ = partial; NR = not clearly reported. For cue/CF supervision, ○* denotes confounder-based supervision and ○** denotes debiased-cue supervision.

label-only meme datasets, MEMECF explicitly focuses on memes whose labels are driven by small but decisive textual, visual, or joint image-text details. Each meme is paired with cue-focused evidence annotations that describe the pivotal detail supporting the classification decision.

Building on this cue-focused formulation, we propose MemePIVOT, an evidential local-global architecture for multimodal meme classification. MemePIVOT is designed around the observation that meme labels are often governed by a small set of pivotal cues. Rather than relying only on global image-text features, MemePIVOT explicitly learns local evidence by aligning textual tokens with image patches. It uses unbalanced optimal transport to identify fine-grained word–patch correspondences, constructs cross-modal evidence from these alignments, and encodes the resulting interactions with a local Transformer. In parallel, a global branch captures holistic image-text context. The local and global branches are then combined through an evidential fusion module that models confidence, uncertainty, and disagreement. This allows MemePIVOT to jointly learn about the overall meme context and the pivotal fine-grained evidence that determines the final prediction.

We evaluate MemePIVOT on three meme classification benchmarks: HarMeme, PrideMM, and

our proposed MemeCF dataset. Across these datasets, MemePIVOT consistently improves over strong multimodal baselines and benchmark vision-language models. On HarMeme, MemePIVOT achieves 83.70 F1 / 92.92 AUROC; on PrideMM, it achieves 76.98 F1 / 84.05 AUROC; and on MemeCF, it achieves 76.88 F1 / 78.20 AUROC for Hate, 87.92 F1 / 94.33 AUROC for Harm, and 82.18 F1 / 85.16 AUROC for Sarcasm. Beyond standard in-dataset evaluation, we conduct cross-dataset experiments, compare against benchmark VLMs, and perform ablation studies to measure the impact of local evidence alignment, global context modeling, and evidential fusion. Our results show that modeling pivotal fine-grained evidence is critical for reliable meme classification, especially when the decisive cues occupy only a small portion of the image-text input.

Our contributions are fourfold:

Contribution 1 — Cue-focused benchmark.

We introduce MEMECF, a cue-focused benchmark of 9,895 memes for fine-grained meme classification across hate, harm, and sarcasm. Unlike label-only meme datasets, MEMECF focuses on examples whose labels are driven by small but decisive textual, visual, or joint image-text details.

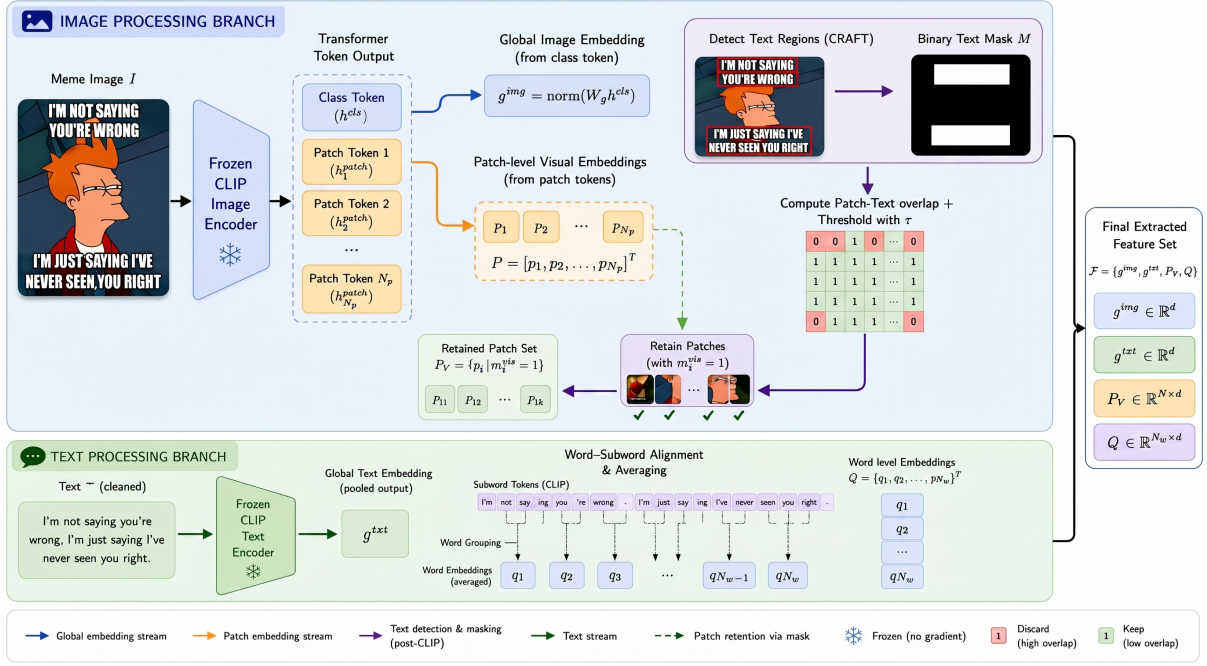


Figure 1: Feature extraction step of MemePIVOT: frozen CLIP encoders produce global, patch-level, and word-level features with text-mask-guided patch retention.

Contribution 2 — Fine-grained evidence annotations. We provide cue-focused evidence annotations that describe the pivotal details supporting each classification decision. These annotations enable fine-grained dataset analysis and evidence-level evaluation beyond image-level labels, while classification models use only the meme image and extracted text as input.

Contribution 3 — Pivotal evidence model. We propose MEMEPIVOT, an evidential local-global architecture for multimodal meme classification. MEMEPIVOT uses unbalanced optimal transport to align textual tokens with image patches, encodes the resulting local cross-modal interactions with a Transformer, and fuses local and global evidence under uncertainty.

Contribution 4 — Comprehensive evaluation. We evaluate MEMEPIVOT across HarMeme, PrideMM, and MEMECF, including in-dataset performance, cross-dataset transfer, architectural ablations, and comparisons with benchmark vision-language models.

2 Related Work

Meme Benchmarks and Cue Evidence. Prior work has established meme classification as a challenging multimodal problem, with benchmarks spanning sentiment, emotion, offensiveness, hate,

harm, cyberbullying, and sarcasm. MEMOTION and MULTIOFF formalized meme analysis beyond ordinary image or text classification (Sharma et al., 2020; Suryawanshi et al., 2020), while HATEFUL MEMES emphasized joint visual-textual reasoning through benign confounders (Kiela et al., 2020). Later datasets broadened the space to harmfulness, target identification, code-mixed cyberbullying, LGBTQ+ hate, fine-grained harm categories, and multimodal sarcasm (Pramanick et al., 2021a; Maity et al., 2022; Shah et al., 2024; El-amrany et al., 2025; Qin et al., 2023; Yue et al., 2024). In parallel, work on rationales and explanations has shown that final-label accuracy alone does not fully characterize model behavior: ERASER evaluates textual evidence spans (DeYoung et al., 2020), VCR studies vision-language rationales (Zellers et al., 2019), and meme-specific work has explored target-aware harmful meme detection and natural-language explanations (Pramanick et al., 2021a; Sharma et al., 2022; Hee et al., 2023). These efforts provide important label-, target-, and rationale-level supervision, but they are not primarily organized around the specific setting where a small word, symbol, object, gesture, or image-text mismatch is sufficient to change the meme’s interpretation.

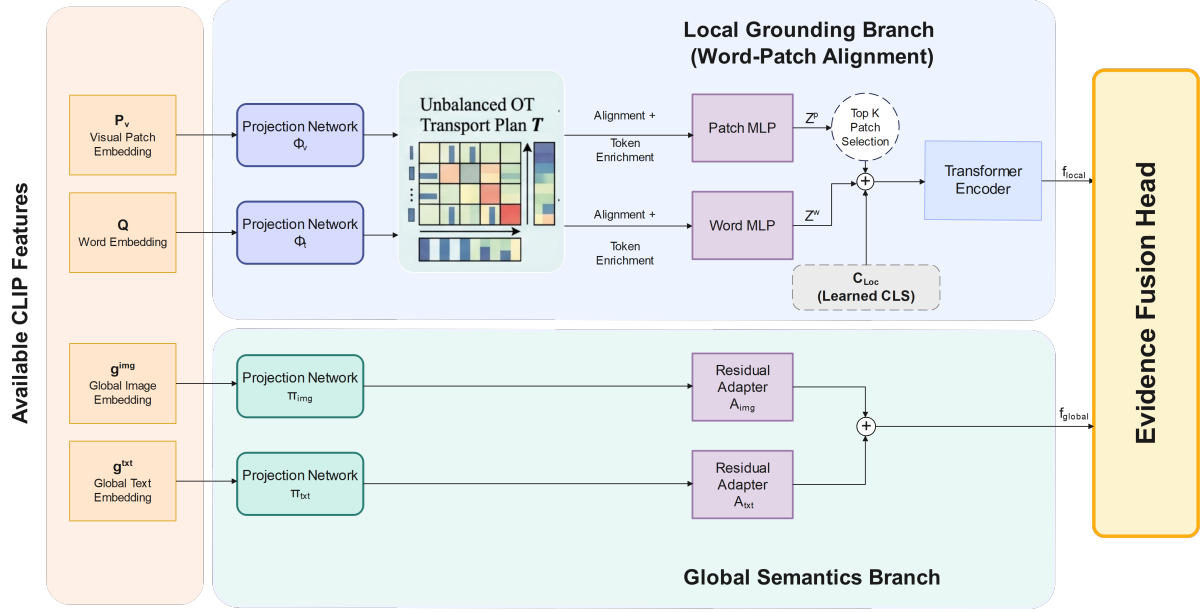


Figure 2: MemePIVOT architecture at a glance: extracted CLIP features are processed through local word–patch grounding and global semantic branches, then passed to the evidence-based fusion head for final classification.

Vision-Language Alignment and Uncertainty.

Large vision-language representations such as CLIP have become standard backbones for multimodal meme understanding (Radford et al., 2021). Meme-specific models adapt these representations through cross-modal feature interaction, lightweight CLIP-based modules, retrieval-guided contrastive learning, prompting, textual inversion, visually grounded context, or semantic tags (Kumar and Nandakumar, 2022; Shah et al., 2024; Mei et al., 2024; Cao et al., 2022; Burbi et al., 2023; Koushik et al., 2026; Sharma and Patil, 2026a). Beyond individual architectures, controlled studies have systematically examined encoder choice, fusion strategy, and modality reliance in multimodal meme classification (Sharma and Patil, 2026b). While these approaches show the strength of pre-trained vision-language features, global embeddings and generated summaries can blur the contribution of small local details when most of the meme provides context and only a small cue determines the label. Local cross-modal alignment offers a complementary direction: attention, graph reasoning, and feature interaction have been used to connect visual and textual information (Schifanella et al., 2016; Cai et al., 2019; Liang et al., 2021, 2022), while optimal transport provides a soft matching formulation across feature sets (Cuturi, 2013). In text-embedded images, unbalanced optimal transport is particularly relevant because OCR tokens may be noisy, visual patches may be irrele-

vant, and not every word or region should be forced to align (Pham et al., 2020). Since local and global evidence may also disagree, evidential learning provides a way to represent both predictions and uncertainty (Sensoy et al., 2018). Together, these lines of work motivate meme classifiers that combine holistic image–text context with fine-grained local interactions and uncertainty-aware fusion.

3 Methodology

3.1 Frozen CLIP Feature Extraction

Given a meme image I and its text T , we extract multimodal representations using frozen CLIP encoders. The CLIP vision transformer represents the image as one class token and N_p patch tokens. Instead of using only the default pooled image embedding, we retain the hidden token sequence before final pooling, $H^{\text{img}} = [h^{\text{cls}}; h_1^{\text{patch}}; \dots; h_{N_p}^{\text{patch}}]$. The class token gives the global image feature $g_{\text{img}} = \text{norm}(W_v h^{\text{cls}})$, while each patch token is projected into the CLIP embedding space as $p_j = \text{norm}(W_v h_j^{\text{patch}})$, forming $P = [p_1, \dots, p_{N_p}]^T$. Thus, g_{img} preserves holistic visual semantics, whereas P provides local visual evidence from CLIP’s internal patch-token states.

To reduce the effect of overlaid meme text on the local visual stream, we use CRAFT to detect text regions and rasterize them into a binary mask M . This mask is not applied to the input image and therefore does not alter the frozen CLIP encoding.

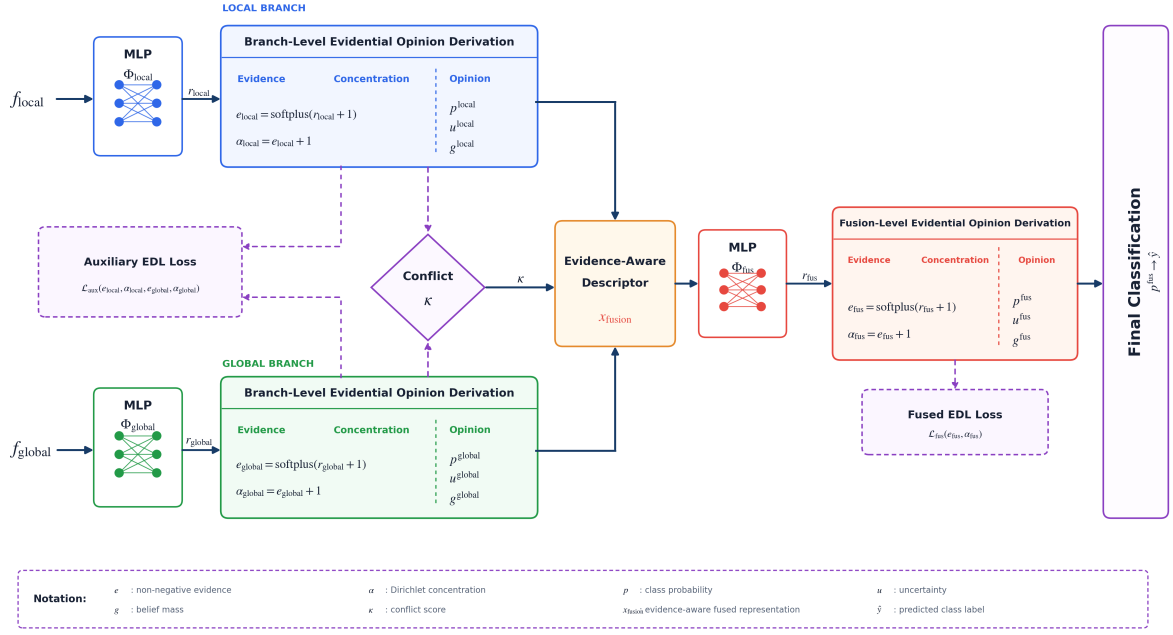


Figure 3: MemePIVOT evidence-based fusion head: local and global evidential opinions are conflict-aware fused to produce uncertainty-calibrated class predictions.

It is used only after feature extraction to measure text overlap for each CLIP patch. Let Ω_j be the image region of the j -th patch. Its text coverage is $\rho_j = |M \cap \Omega_j|/|\Omega_j|$, and the retention mask is $m_j^{\text{vis}} = \mathbf{1}[\rho_j \leq \tau]$, where τ is the text-overlap threshold. The retained local visual set is $P_V = \{p_j \mid m_j^{\text{vis}} = 1\}$. This keeps the global image feature unchanged while allowing the local branch to ignore text-dominated patches.

For the text modality, the cleaned text T is encoded using the frozen CLIP text encoder. The pooled output is used as the global text embedding g_{txt} . Since CLIP uses subword tokenization, each word is aligned to its subword positions, and the corresponding projected hidden states are averaged to obtain word-level embeddings $Q = [q_1, \dots, q_{N_w}]^\top$. The final feature set is $\mathcal{F} = \{g_{\text{img}}, g_{\text{txt}}, P_V, Q\}$. The global branch uses g_{img} and g_{txt} for holistic meme semantics, while the local branch uses P_V and Q for fine-grained word-region grounding.

3.2 Local-Global Multimodal Evidence Modeling

3.2.1 Local Grounding Branch

The local branch models fine-grained alignment between words and visual regions. Given P_V and Q , the visual and textual tokens are projected into a shared transport space of dimension $d_{\text{ot}} = 384$

using the projection networks Φ_v and Φ_t : $\tilde{p}_j = \text{norm}(\Phi_v(p_j))$ and $\tilde{q}_i = \text{norm}(\Phi_t(q_i))$. Their cross-modal cost is $D_{ij} = 1 - \tilde{q}_i^\top \tilde{p}_j$, so semantically compatible word-patch pairs have lower cost.

Rather than forcing every word and patch to match, we solve an unbalanced optimal transport problem with dustbin states. This allows irrelevant words or patches to remain weakly matched, reducing spurious alignments. Let $\mathbf{T}$ denote the resulting transport plan, where bold notation distinguishes it from the input text T . For each word, the aligned visual summary is $\hat{p}_i = \sum_j \mathbf{T}_{ij} \tilde{p}_j / (\sum_j \mathbf{T}_{ij} + \epsilon)$, and for each patch, the aligned textual summary is $\hat{q}_j = \sum_i \mathbf{T}_{ij} \tilde{q}_i / (\sum_i \mathbf{T}_{ij} + \epsilon)$. The dustbin terms $d_i^{(w)}$ and $d_j^{(p)}$ measure unmatched word-side and patch-side evidence.

We form enriched local descriptors $\bar{z}_i^w = [\tilde{q}_i; \hat{p}_i; \tilde{q}_i - \hat{p}_i; d_i^{(w)}]$ and $\bar{z}_j^p = [\tilde{p}_j; \hat{q}_j; \tilde{p}_j - \hat{q}_j; d_j^{(p)}]$, each in $\mathbb{R}^{3d_{\text{ot}}+1}$. The Word MLP and Patch MLP map these descriptors into a local hidden dimension $d_{\text{local}} = 256$, producing Z^w and Z^p . To suppress noisy visual evidence, each patch is scored by its transported mass $s_j = \sum_i \mathbf{T}_{ij}$, and only the top K patch tokens are retained as $Z_{\text{top}K}^p$. The word tokens, selected patch tokens, and learned local classification token C_{Loc} are then concatenated and passed through a lightweight transformer:

$f_{\text{local}} = \text{Transformer}_{\text{loc}}(C_{\text{Loc}} \oplus Z^w \oplus Z_{\text{top}K}^p)_0$, where $\oplus$ denotes token concatenation. This branch explicitly grounds text in relevant image regions, handles unmatched evidence through dustbins, and aggregates the most informative local cues.

3.2.2 Global Semantic Branch

The global branch preserves holistic meme-level semantics that may be lost when using only sparse local alignments. It operates on the frozen CLIP global embeddings g_{img} and g_{txt} . The global projection networks π_{img} and π_{txt} map them into a shared latent dimension $d_g = 512$, giving $u_{\text{img}} = \pi_{\text{img}}(g_{\text{img}})$ and $u_{\text{txt}} = \pi_{\text{txt}}(g_{\text{txt}})$. These features are refined using residual adapters $A(x) = x + W_2\sigma(W_1 \text{LN}(x))$, where $\sigma(\cdot)$ is GELU and $\text{LN}(\cdot)$ is layer normalization. The adapted features are $h_{\text{img}} = A_{\text{img}}(u_{\text{img}})$ and $h_{\text{txt}} = A_{\text{txt}}(u_{\text{txt}})$, and the global representation is $f_{\text{global}} = h_{\text{img}} \oplus h_{\text{txt}} \in \mathbb{R}^{2d_g}$, where $\oplus$ denotes feature concatenation. Unlike the local branch, this pathway does not perform patch pruning or token-level alignment; it retains broad visual context, overall text–image compatibility, and scene-level semantic cues.

Together, the two branches provide complementary evidence: f_{local} captures grounded word–region interactions, while f_{global} preserves high-level multimodal meaning.

3.2.3 Evidence-Based Fusion Head

After obtaining f_{local} and f_{global} , the model fuses them using evidential opinions instead of directly averaging logits. For each branch $m \in \{\text{local}, \text{global}\}$, the branch MLP Φ_m produces raw scores $r_m = \Phi_m(f_m)$, which are converted into non-negative evidence $e_m = \text{softplus}(r_m)$ and Dirichlet parameters $\alpha_m = e_m + 1$. Let N_c denote the number of classes and $S_m = \sum_{c=1}^{N_c} \alpha_{m,c}$. The branch predictive probability, belief mass, and uncertainty are $p_c^m = \alpha_{m,c}/S_m$, $g_c^m = e_{m,c}/S_m$, and $u^m = N_c/S_m$, respectively. Higher evidence therefore produces lower uncertainty, while weak evidence gives a flatter and less confident opinion.

The fusion module receives the evidence-aware descriptor

$$x_{\text{fusion}} = [e_{\text{local}}; e_{\text{global}}; g^{\text{local}}; g^{\text{global}}; u^{\text{local}}; u^{\text{global}}; \kappa],$$

where the conflict term $\kappa = (p_+^{\text{local}} - p_+^{\text{global}})^2$ measures disagreement between the local and global positive-class probabilities. This helps the model

Task	All 0/1	Train 0/1	Val 0/1	Test 0/1
Harm	1397 / 1621	1045 / 1218	139 / 163	213 / 240
Hate	1500 / 1789	1127 / 1340	152 / 177	221 / 272
Sarcasm	1639 / 1949	1233 / 1458	164 / 195	242 / 296

Table 2: MemeCF statistics over 9,895 memes. Label 0/1 denotes negative/positive classes for each task: Harm, Hate, and Sarcasm.

distinguish confident agreement from confident disagreement.

The fusion MLP Φ_{fus} maps x_{fusion} to raw fused scores $r_{\text{fus}} = \Phi_{\text{fus}}(x_{\text{fusion}})$, which are converted into fused evidence $e_{\text{fus}} = \text{softplus}(r_{\text{fus}})$ and fused Dirichlet parameters $\alpha_{\text{fus}} = e_{\text{fus}} + 1$. With $S_{\text{fus}} = \sum_{c=1}^{N_c} \alpha_{\text{fus},c}$, the final predictive probability, belief mass, and uncertainty are $p_c^{\text{fus}} = \alpha_{\text{fus},c}/S_{\text{fus}}$, $g_c^{\text{fus}} = e_{\text{fus},c}/S_{\text{fus}}$, and $u^{\text{fus}} = N_c/S_{\text{fus}}$. The final prediction is $\hat{y} = \arg \max_c p_c^{\text{fus}}$.

This fusion head acts as an evidence-aware arbitration module. It can rely more on the local branch when word–region grounding is confident, rely more on the global branch when holistic meme semantics are clearer, or reduce confidence when the branches are uncertain or contradictory. The model is trained by supervising the fused output and both branch-level opinions. Let

$$\begin{aligned}\mathcal{L}_f &= \mathcal{L}_{\text{EDL}}(e_{\text{fus}}, \alpha_{\text{fus}}, y), \\ \mathcal{L}_{\text{loc}} &= \mathcal{L}_{\text{EDL}}(e_{\text{local}}, \alpha_{\text{local}}, y), \\ \mathcal{L}_{\text{glob}} &= \mathcal{L}_{\text{EDL}}(e_{\text{global}}, \alpha_{\text{global}}, y).\end{aligned}$$

The total loss is

$$\begin{aligned}\mathcal{L} &= \mathcal{L}_f + \lambda_{\text{local}}\mathcal{L}_{\text{loc}} + \lambda_{\text{global}}\mathcal{L}_{\text{glob}} \\ &\quad + \lambda_{\text{uot}}\mathcal{L}_{\text{uot}} + \lambda_{\text{dust}}\mathcal{L}_{\text{dust}}.\end{aligned}$$

The auxiliary branch losses keep both local and global representations discriminative, while the transport and dustbin losses encourage stable alignment and controlled unmatched evidence. Overall, the final decision is formed by reconciling fine-grained grounding cues with holistic multimodal context on a per-sample basis.

4 Experimental Details and Result Analysis

4.1 MemeCF Dataset

We introduce **MemeCF**, a counterfactual meme dataset for three binary tasks: harm, hate, and sarcasm. It contains 9,895 memes sourced from Multi-Bully, MMSD 2.0, PrideMM, HMC, HarMeme,

and GuardHarMem (Maity et al., 2022; Qin et al., 2023; Shah et al., 2024; Kiela et al., 2020; Pramanick et al., 2021a; El-amrany et al., 2025). Each instance includes an image, meme text, task label, modality-level counterfactual evidence indicating whether the label-changing cue lies in the text, image, or both, and supporting reasoning. Candidate counterfactuals were filtered through a LLM-as-judge protocol followed by human review. Further details, including image-selection criteria, examples, and extended distributions, are provided in the Supplementary Material.

Table 2 reports the task-wise label distribution. Each task is split into approximately 75%/10%/15% train, validation, and test partitions while preserving label balance.

4.2 Dataset Configuration

We evaluate MemePIVOT on three binary multi-modal meme benchmarks: HarMeme, PrideMM, and MemeCF, covering harmful, hateful, and sarcasm-related meme understanding. Each dataset consists of image-text meme pairs (I, T_o) with task-specific labels.

HarMeme (Pramanick et al., 2021a) contains 3,544 COVID-19 memes labeled harmful or harmless and is evaluated using its official splits. PrideMM (Shah et al., 2024) includes 5,063 LGBTQ+ memes, split into 4,328/228/507 train, validation, and test examples. MemeCF, our proposed counterfactual meme dataset, contains 9,895 memes across harm, hate, and sarcasm, with each task split into approximately 75%/10%/15% train, validation, and test partitions while preserving label distributions, as shown in Table 2. Models are trained on the training split, tuned on validation data, and evaluated on the held-out test split. Additional implementation details are provided in the Supplementary Material.

4.3 Experimental Results

As shown in Tables 3 and 4, MemePIVOT achieves the best performance on five of six HarMeme/PrideMM metrics and eight of nine MemeCF metrics. On HarMeme, it improves over the strongest non-MemePIVOT result by 2.31 Acc., 1.61 AUROC, and 2.00 F1 points. On PrideMM, it improves Acc. and F1 by 1.78 and 1.81 points, respectively, while its AUROC of 84.05 is 0.02 points below CLIPAdapter’s 84.07. On MemeCF, MemePIVOT leads all three Hate metrics, all three Harm metrics, and Sarcasm Acc. and F1; Meme-

CLIP obtains the best Sarcasm AUROC. These results show strong and consistent performance across datasets, with the remaining difference confined to one metric.

4.3.1 Cross-Dataset Evaluation

We evaluate MemePIVOT alongside ISSUES, MemeCLIP, and HarMDetect under a controlled cross-dataset setting. Each model is trained on the PrideMM (Shah et al., 2024) hate task and evaluated on the HMC (Kiela et al., 2020) test set, preserving the same binary hate-classification task definition while changing the dataset distribution. As shown in Table 5, MemePIVOT achieves the strongest performance across all three metrics, outperforming the strongest comparison baseline, ISSUES, by 3.10 Acc., 2.66 AUROC, and 4.20 F1 percentage points. This indicates that the learned pivotal-evidence representation transfers when evaluated under cross-dataset/out-of-domain setting

5 Ablation Studies

In this section, we present an ablation study evaluating the contribution of each proposed model component to overall performance. In addition, the supplementary material includes an ablation study benchmarking state-of-the-art LLMs on meme classification using the HarMeme test set. The results highlight two key findings: (i) trained meme classification models continue to outperform state-of-the-art LLMs as of 2026, and (ii) state-of-the-art LLMs are substantially more expensive to deploy, making their large-scale use for internet-scale meme moderation infeasible.

5.1 Impact of each Component on Overall Model Performance

We perform a cumulative ablation to isolate the effect of each component on model performance using the HarMeme dataset, as shown in Table 6. Adding the local branch raises Acc. from 84.56 to 85.86 and F1 from 78.12 to 79.48, while AUROC changes from 91.38 to 91.10. Adding the evidential head further improves Acc., AUROC, and F1 to 87.23, 92.08, and 81.52, respectively. Adding the optimal-transport loss produces the strongest configuration, reaching 88.04 Acc., 92.92 AUROC, and 83.70 F1, corresponding to gains of 0.81, 0.84, and 2.18 points over the preceding configuration. Overall, the full model improves over the global-only variant by 3.48 Acc., 1.54 AUROC, and 5.58

Table 3: Performance comparison across HarMeme and PrideMM using mean $\pm$ standard deviation. Single-run VLM results marked by $\dagger$.

Model	HarMeme			PrideMM		
	Acc.	AUROC	F1	Acc.	AUROC	F1
<i>Text-only models</i>						
BERT (Devlin et al., 2019)	75.83 $\pm$ 1.25	82.57 $\pm$ 1.06	68.31 $\pm$ 0.86	67.85 $\pm$ 1.41	78.02 $\pm$ 0.59	71.58 $\pm$ 1.01
CLIPTextOnly (Radford et al., 2021)	75.81 $\pm$ 1.43	80.01 $\pm$ 1.58	59.26 $\pm$ 3.62	69.87 $\pm$ 2.10	75.86 $\pm$ 0.73	69.86 $\pm$ 1.13
<i>Image-only models</i>						
CLIPImgOnly (Radford et al., 2021)	74.16 $\pm$ 1.97	82.92 $\pm$ 0.83	67.98 $\pm$ 2.43	62.65 $\pm$ 1.93	68.73 $\pm$ 2.01	60.01 $\pm$ 3.47
<i>Vision-language models</i>						
LLaVA-1.5-7B $\dagger$ (Liu et al., 2024)	72.88	66.31	43.53	59.96	45.27	55.19
Qwen2.5-VL $\dagger$ (Bai et al., 2025b)	83.33	90.13	78.39	75.12	82.21	75.09
<i>Other multimodal models</i>						
CLIPAdapter (Gao et al., 2024)	81.93 $\pm$ 2.12	91.14 $\pm$ 0.71	78.03 $\pm$ 1.34	74.89 $\pm$ 1.20	84.07 $\pm$ 0.53	74.92 $\pm$ 1.23
M2CLIP_SEmoNet (Dhiman et al., 2026)	81.87 $\pm$ 0.92	89.91 $\pm$ 1.53	73.33 $\pm$ 2.08	74.27 $\pm$ 1.58	82.11 $\pm$ 1.01	75.17 $\pm$ 1.60
DORA (Hossain et al., 2024)	82.81 $\pm$ 1.39	88.02 $\pm$ 1.84	76.01 $\pm$ 2.15	73.76 $\pm$ 1.29	81.74 $\pm$ 0.53	73.11 $\pm$ 1.82
HateCLIPPER (Kumar and Nandakumar, 2022)	85.73 $\pm$ 0.61	90.27 $\pm$ 0.21	80.84 $\pm$ 0.60	74.81 $\pm$ 0.61	82.54 $\pm$ 0.49	74.39 $\pm$ 0.47
HarMDetect (El-amrany et al., 2025)	84.01 $\pm$ 0.25	90.24 $\pm$ 1.27	78.21 $\pm$ 1.02	73.51 $\pm$ 0.12	81.02 $\pm$ 0.20	73.15 $\pm$ 0.53
MemeCLIP (Shah et al., 2024)	84.75 $\pm$ 0.28	91.31 $\pm$ 0.35	81.70 $\pm$ 0.74	75.11 $\pm$ 0.41	82.92 $\pm$ 0.53	73.67 $\pm$ 0.60
MemePIVOT	88.04 $\pm$ 0.59	92.92 $\pm$ 0.40	83.70 $\pm$ 0.48	76.90 $\pm$ 0.83	84.05 $\pm$ 0.11	76.98 $\pm$ 0.73

Table 4: Performance comparison across MemeCF tasks. Multi-seed results are reported as mean $\pm$ standard deviation; single-run zero-shot VLM baselines are marked by $\dagger$.

Model	MemeCF-Hate			MemeCF-Harm			MemeCF-Sarcasm		
	Acc.	AUROC	F1	Acc.	AUROC	F1	Acc.	AUROC	F1
<i>Text-only models</i>									
BERT (Devlin et al., 2019)	63.08 $\pm$ 3.26	69.44 $\pm$ 2.13	69.78 $\pm$ 3.36	73.14 $\pm$ 0.13	81.34 $\pm$ 0.14	76.82 $\pm$ 0.31	68.59 $\pm$ 2.81	74.68 $\pm$ 3.14	73.72 $\pm$ 1.73
CLIPTextOnly (Radford et al., 2021)	59.43 $\pm$ 3.96	63.12 $\pm$ 8.37	71.08 $\pm$ 0.89	62.33 $\pm$ 3.17	72.51 $\pm$ 0.79	71.63 $\pm$ 0.36	64.13 $\pm$ 5.70	70.71 $\pm$ 0.48	74.52 $\pm$ 2.04
<i>Image-only models</i>									
CLIPImgOnly (Radford et al., 2021)	58.62 $\pm$ 2.82	62.85 $\pm$ 2.24	70.32 $\pm$ 1.08	71.30 $\pm$ 3.11	79.57 $\pm$ 1.07	77.30 $\pm$ 1.10	64.62 $\pm$ 2.45	69.53 $\pm$ 1.98	72.10 $\pm$ 2.86
<i>VLM baselines</i>									
Gemma 3 Vision 12B $\dagger$ (Gemma Team, 2025)	70.39	75.32	74.30	62.69	62.81	69.77	61.15	67.07	73.17
LLaVA-1.5-7B $\dagger$ (Liu et al., 2024)	58.62	68.90	52.56	53.64	52.65	59.92	57.62	58.67	71.21
Qwen3-VL-8B $\dagger$ (Bai et al., 2025a)	68.76	75.10	74.84	60.26	64.87	68.31	63.75	69.37	74.77
Phi-4 Multimodal $\dagger$ (Abouelenin et al., 2025)	66.73	73.52	69.74	58.50	62.26	59.48	66.91	72.10	72.53
<i>Other multimodal models</i>									
CLIPAdapter (Gao et al., 2024)	68.29 $\pm$ 1.25	74.06 $\pm$ 0.75	73.25 $\pm$ 1.71	82.27 $\pm$ 1.00	90.11 $\pm$ 0.83	83.79 $\pm$ 0.88	77.20 $\pm$ 1.67	83.92 $\pm$ 0.85	80.45 $\pm$ 2.72
M2CLIP_SEmoNet (Dhiman et al., 2026)	68.36 $\pm$ 2.34	74.97 $\pm$ 1.12	72.74 $\pm$ 1.00	84.91 $\pm$ 0.48	92.93 $\pm$ 0.29	86.13 $\pm$ 0.40	71.50 $\pm$ 0.09	77.85 $\pm$ 0.98	75.91 $\pm$ 1.15
HarMDetect (El-amrany et al., 2025)	68.22 $\pm$ 1.84	72.57 $\pm$ 1.11	73.51 $\pm$ 1.38	83.91 $\pm$ 1.24	92.73 $\pm$ 0.75	85.31 $\pm$ 1.31	74.78 $\pm$ 0.76	82.00 $\pm$ 0.15	78.15 $\pm$ 0.63
DORA (Hossain et al., 2024)	65.86 $\pm$ 1.15	71.42 $\pm$ 1.29	73.12 $\pm$ 0.57	82.78 $\pm$ 1.81	91.90 $\pm$ 1.79	84.67 $\pm$ 1.47	73.36 $\pm$ 0.65	79.92 $\pm$ 0.57	77.55 $\pm$ 1.68
MemeCLIP (Shah et al., 2024)	66.60 $\pm$ 0.23	74.29 $\pm$ 0.64	66.10 $\pm$ 0.10	83.37 $\pm$ 0.67	92.61 $\pm$ 0.23	83.27 $\pm$ 0.67	76.83 $\pm$ 0.88	85.61 $\pm$ 0.31	76.08 $\pm$ 0.97
HateCLIPPER (Kumar and Nandakumar, 2022)	71.08 $\pm$ 0.62	76.94 $\pm$ 0.62	75.52 $\pm$ 1.38	84.91 $\pm$ 0.21	92.98 $\pm$ 0.12	86.09 $\pm$ 0.23	77.06 $\pm$ 1.09	81.22 $\pm$ 0.96	80.57 $\pm$ 0.95
ISSUES (Burbi et al., 2023)	70.59 $\pm$ 1.05	76.54 $\pm$ 0.30	74.47 $\pm$ 1.39	82.49 $\pm$ 0.92	90.93 $\pm$ 1.20	83.85 $\pm$ 0.80	77.20 $\pm$ 1.55	85.05 $\pm$ 0.83	80.52 $\pm$ 1.47
MemePIVOT	71.38 $\pm$ 0.23	78.20 $\pm$ 0.57	76.88 $\pm$ 1.02	86.92 $\pm$ 1.29	94.33 $\pm$ 0.22	87.92 $\pm$ 1.22	79.18 $\pm$ 0.66	85.16 $\pm$ 0.89	82.18 $\pm$ 0.71

Table 5: Cross-dataset hate-classification evaluation under dataset shift, with models trained on PrideMM and evaluated on HMC using the same task definition.

Model	Acc.	AUROC	F1
ISSUES (Burbi et al., 2023)	0.5796	0.6021	0.5938
MemeCLIP (Shah et al., 2024)	0.5776	0.5994	0.5879
HarMDetect (El-amrany et al., 2025)	0.5455	0.5861	0.5709
MemePIVOT	0.6106	0.6287	0.6358

F1 points.

6 Conclusion

We introduced MemeCF and MemePIVOT for cue-focused multimodal meme classification. MemeCF focuses on harm, hate, and sarcasm cases where the label depends on small but decisive textual, visual, or joint image-text cues, and provides evidence

Table 6: Cumulative ablation study on HarMeme showcasing impact of each component. G denotes the global branch, L denotes the local branch, E denotes the evidential head, and OT denotes the optimal-transport loss. Components are added progressively from top to bottom. Results are reported as mean $\pm$ standard deviation.

G	L	E	OT	Acc.	AUROC	F1
✓	–	–	–	84.56 $\pm$ 0.38	91.38 $\pm$ 0.98	78.12 $\pm$ 1.84
✓	✓	–	–	85.86 $\pm$ 1.12	91.10 $\pm$ 0.82	79.48 $\pm$ 0.54
✓	✓	✓	–	87.23 $\pm$ 0.71	92.08 $\pm$ 0.43	81.52 $\pm$ 1.18
✓	✓	✓	✓	88.04 $\pm$ 0.59	92.92 $\pm$ 0.40	83.70 $\pm$ 0.48

annotations describing the pivotal details behind each decision. MemePIVOT builds on this motivation by combining local word-patch grounding with global CLIP-based meme semantics, followed by evidential fusion to account for uncertainty and disagreement between local and global signals.

The experiments show that MemePIVOT is effective across the evaluated benchmarks. It achieves 88.04 accuracy, 92.92 AUROC, and 83.70 F1 on HarMeme, and 76.90 accuracy, 84.05 AUROC, and 76.98 F1 on PrideMM. On MemeCF, it obtains strong performance across all three sub-tasks: 71.38 accuracy, 78.20 AUROC, and 76.88 F1 for hate; 86.92 accuracy, 94.33 AUROC, and 87.92 F1 for harm; and 79.18 accuracy, 85.16 AUROC, and 82.18 F1 for sarcasm. The cross-dataset setting further indicates that the learned pivotal-evidence representation transfers well under domain shift, and the ablation study confirms the value of local grounding, evidential fusion, and optimal-transport alignment.

Future work can build directly on the evidence annotations in MemeCF by using them not only for dataset analysis, but also for explicit evidence-level supervision and evaluation. Another direction is to extend the cue-focused formulation to broader meme domains, languages, and cultural contexts, where pivotal cues may be more implicit or context-dependent. Finally, given the paper’s finding that general-purpose multimodal LLMs remain costly and inconsistent in zero-shot harmful meme classification, future work can explore more efficient task-specific models that preserve fine-grained cue sensitivity while reducing inference cost.

Limitations

Despite the complementary design of the local branch for subtle cue discovery and the global branch for holistic semantic understanding, the model remains fundamentally constrained by the frozen CLIP encoder used to produce its representations. In particular, culturally grounded humor, symbolism, slang, and historically situated references are only captured to the extent that they are already encoded in the pretrained backbone. As a result, the model may underperform on culturally sensitive memes whose harmful or sarcastic meaning depends on community-specific knowledge, implicit social context, or visual-textual references that are weakly represented in CLIP’s pretraining distribution.

In addition, the evidential outputs should not be interpreted as a guarantee of calibrated reliability. Although the model produces Dirichlet-style evidence, belief, and uncertainty estimates, these quantities are still learned from the training distribution and inherit its biases and coverage gaps.

Under distribution shift, unseen meme formats, unfamiliar cultural references, or ambiguous annotations, the model can still become overconfident despite producing uncertainty scores. Therefore, the uncertainty estimates are better viewed as a useful diagnostic signal rather than a substitute for calibration analysis, robustness testing, or human review.

References

- Abdelrahman Abouelenin, Atabak Ashfaq, Adam Atkinson, Hany Awadalla, Nguyen Bach, Jianmin Bao, Alon Benhaim, Martin Cai, Vishrav Chaudhary, Congcong Chen, Dong Chen, Dongdong Chen, Junkun Chen, Weizhu Chen, Yen-Chun Chen, Yi-ling Chen, Qi Dai, Xiyang Dai, Ruchao Fan, and 55 others. 2025. [Phi-4-Mini technical report: Compact yet powerful multimodal language models via mixture-of-LoRAs](#). *CoRR*, abs/2503.01743.
- Anthropic. 2026. [Introducing Claude Opus 4.6](#).
- Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhi-fang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, and 45 others. 2025a. [Qwen3-VL technical report](#). *CoRR*, abs/2511.21631.
- Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibao Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Ming-Hsuan Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, and 8 others. 2025b. [Qwen2.5-VL technical report](#). *CoRR*, abs/2502.13923.
- Aashish Bhandari, Siddhant B. Shah, Surendrabikram Thapa, Usman Naseem, and Mehwish Nasim. 2023. [CrisisHateMM: Multimodal analysis of directed and undirected hate speech in text-embedded images from russia-ukraine conflict](#). In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 1994–2003.
- Giovanni Burbi, Alberto Baldrati, Lorenzo Agnolucci, Marco Bertini, and Alberto Del Bimbo. 2023. [Mapping memes to words for multimodal hateful meme classification](#). In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*, pages 2832–2836.
- Yitao Cai, Huiyu Cai, and Xiaojun Wan. 2019. [Multimodal sarcasm detection in Twitter with hierarchical fusion model](#). In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2506–2515, Florence, Italy. Association for Computational Linguistics.

- Rui Cao, Roy Ka-Wei Lee, Wen-Haw Chong, and Jing Jiang. 2022. [Prompting for multimodal hateful meme classification](#). In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 321–332, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.
- Marco Cuturi. 2013. [Sinkhorn distances: Lightspeed computation of optimal transport](#). In *Advances in Neural Information Processing Systems*, volume 26.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. [BERT: Pre-training of deep bidirectional transformers for language understanding](#). In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.
- Jay DeYoung, Sarthak Jain, Nazneen Fatema Rajani, Eric Lehman, Caiming Xiong, Richard Socher, and Byron C. Wallace. 2020. [ERASER: A benchmark to evaluate rationalized NLP models](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4443–4458, Online. Association for Computational Linguistics.
- Chhavi Dhiman, Saihajpreet Singh, and Satyam Mohan. 2026. [Enhancing memes analysis with multi-task multimodal M2CLIP-SEmoNet architecture](#). *Signal Processing: Image Communication*, 144:117526.
- Samir El-amrany, Salima Lamsiyah, Matthias R. Brust, and Pascal Bouvry. 2025. [GuardHarMem and HarMDetect: a multimodal dataset and benchmark model for fine-grained harmful meme classification](#). *Social Network Analysis and Mining*, 15:63.
- Peng Gao, Shijie Geng, Renrui Zhang, Teli Ma, Rongyao Fang, Yongfeng Zhang, Hongsheng Li, and Yu Qiao. 2024. [CLIP-adapter: Better vision-language models with feature adapters](#). *International Journal of Computer Vision*, 132(2):581–595.
- Gemma Team. 2025. [Gemma 3 technical report](#). *CoRR*, abs/2503.19786.
- Google. 2025. [Gemini 3 Flash: Frontier intelligence built for speed](#).
- Google. 2026. [Gemini 3.1 Pro: A smarter model for your most complex tasks](#).
- Ming Shan Hee, Wen-Haw Chong, and Roy Ka-Wei Lee. 2023. [Decoding the underlying meaning of multimodal hateful memes](#). In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence, IJCAI-23*, pages 5995–6003. International Joint Conferences on Artificial Intelligence Organization. AI for Good.
- Eftekhari Hossain, Omar Sharif, Mohammed Moshiri Hoque, and Sarah M. Preum. 2024. [Deciphering hate: Identifying hateful memes and their targets](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*, pages 8347–8359, Bangkok, Thailand. Association for Computational Linguistics.
- 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](#). In *Advances in Neural Information Processing Systems*, volume 33, pages 2611–2624. Curran Associates, Inc.
- Girish A. Koushik, Helen Treharne, Aditya Joshi, and Diptesh Kanojia. 2026. [Trace: Textual relevance augmentation and contextual encoding for multimodal hate detection](#). In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 38763–38771.
- Gokul Karthik Kumar and Karthik Nandakumar. 2022. [Hate-CLIPper: Multimodal hateful meme classification based on cross-modal interaction of CLIP features](#). In *Proceedings of the Second Workshop on NLP for Positive Impact (NLP4PI)*, pages 171–183, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.
- Bin Liang, Chenwei Lou, Xiang Li, Lin Gui, Min Yang, and Ruifeng Xu. 2021. [Multi-modal sarcasm detection with interactive in-modal and cross-modal graphs](#). In *Proceedings of the 29th ACM International Conference on Multimedia, MM '21*, pages 4707–4715, New York, NY, USA. Association for Computing Machinery.
- Bin Liang, Chenwei Lou, Xiang Li, Min Yang, Lin Gui, Yulan He, Wenjie Pei, and Ruifeng Xu. 2022. [Multi-modal sarcasm detection via cross-modal graph convolutional network](#). In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1767–1777, Dublin, Ireland. Association for Computational Linguistics.
- Hongzhan Lin, Ziyang Luo, Bo Wang, Ruichao Yang, and Jing Ma. 2026. [GOAT-Bench: Safety insights to large multimodal models through meme-based social abuse](#). *ACM Transactions on Intelligent Systems and Technology*, 17(4):1–25. Article 79.
- Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2024. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26296–26306.
- Krishanu Maity, Prince Jha, Sriparna Saha, and Pushpak Bhattacharyya. 2022. [A multitask framework for sentiment, emotion and sarcasm aware cyberbullying detection from multi-modal code-mixed memes](#). In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '22*, pages 1739–1749, New York, NY, USA. Association for Computing Machinery.

- Jingbiao Mei, Jinghong Chen, Weizhe Lin, Bill Byrne, and Marcus Tomalin. 2024. [Improving hateful meme detection through retrieval-guided contrastive learning](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5333–5347, Bangkok, Thailand. Association for Computational Linguistics.
- Moonshot AI. 2026. [Kimi K2.6: From code to creation, from one to many](#).
- OpenAI. 2026a. [Introducing GPT-5.4](#).
- OpenAI. 2026b. [Introducing GPT-5.5](#).
- Khiem Pham, Khang Le, Nhat Ho, Tung Pham, and Hung Bui. 2020. [On unbalanced optimal transport: An analysis of Sinkhorn algorithm](#). In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 7673–7682. PMLR.
- Shraman Pramanick, Dimitar Dimitrov, Rituparna Mukherjee, Shivam Sharma, Md. Shad Akhtar, Preslav Nakov, and Tanmoy Chakraborty. 2021a. [Detecting harmful memes and their targets](#). In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 2783–2796, Online. Association for Computational Linguistics.
- Shraman Pramanick, Shivam Sharma, Dimitar Dimitrov, Md. Shad Akhtar, Preslav Nakov, and Tanmoy Chakraborty. 2021b. [MOMENTA: A multimodal framework for detecting harmful memes and their targets](#). In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 4439–4455, Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Libo Qin, Shijue Huang, Qiguang Chen, Chenran Cai, Yudi Zhang, Bin Liang, Wanxiang Che, and Ruifeng Xu. 2023. [MMSD2.0: Towards a reliable multimodal sarcasm detection system](#). In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 10834–10845, Toronto, Canada. Association for Computational Linguistics.
- Jingnong Qu, Liunian Harold Li, Jieyu Zhao, Sunipa Dev, and Kai-Wei Chang. 2022. [DisinfoMeme: A multimodal dataset for detecting meme intentionally spreading out disinformation](#). *Preprint*, arXiv:2205.12617.
- Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. [Learning transferable visual models from natural language supervision](#). In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR.
- Rossano Schifanella, Paloma de Juan, Joel Tetreault, and Liangliang Cao. 2016. [Detecting sarcasm in multimodal social platforms](#). In *Proceedings of the 24th ACM International Conference on Multimedia, MM '16*, pages 1136–1145, New York, NY, USA. Association for Computing Machinery.
- Murat Sensoy, Lance Kaplan, and Melih Kandemir. 2018. [Evidential deep learning to quantify classification uncertainty](#). In *Advances in Neural Information Processing Systems*, volume 31.
- Siddhant Bikram Shah, Shuvam Shiwakoti, Maheep Chaudhary, and Haohan Wang. 2024. [MemeCLIP: Leveraging CLIP representations for multimodal meme classification](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 17320–17332, Miami, Florida, USA. Association for Computational Linguistics.
- Akshit Sharma and Prashant W. Patil. 2026a. [Memetag: Keyword-driven meme classification through tag embedding reconstruction](#). In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 7679–7688.
- Akshit Sharma and Prashant W. Patil. 2026b. What makes multimodal meme classification reliable? a large-scale controlled study of fusion, encoders, and modality reliance. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 7556–7565.
- Chhavi Sharma, Deepesh Bhageria, William Scott, Srinivas PYKL, Amitava Das, Tanmoy Chakraborty, Viswanath Pulabaigari, and Björn Gambäck. 2020. [SemEval-2020 task 8: Memotion analysis- the visuo-lingual metaphor!](#) In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 759–773, Barcelona (online). International Committee for Computational Linguistics.
- Shivam Sharma, Md Shad Akhtar, Preslav Nakov, and Tanmoy Chakraborty. 2022. [DISARM: Detecting the victims targeted by harmful memes](#). In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 1572–1588, Seattle, United States. Association for Computational Linguistics.
- Shardul Suryawanshi, Bharathi Raja Chakravarthi, Michael Arcan, and Paul Buitelaar. 2020. [Multimodal meme dataset \(MultiOFF\) for identifying offensive content in image and text](#). In *Proceedings of the Second Workshop on Trolling, Aggression and Cyberbullying*, pages 32–41, Marseille, France. European Language Resources Association (ELRA).
- Kohtaro Tanaka, Hiroaki Yamane, Yusuke Mori, Yusuke Mukuta, and Tatsuya Harada. 2022. [Learning to evaluate humor in memes based on the incongruity theory](#). In *Proceedings of the Second Workshop on When Creative AI Meets Conversational AI*, pages 81–93, Gyeongju, Republic of Korea. Association for Computational Linguistics.
- Tan Yue, Xuzhao Shi, Rui Mao, Zonghai Hu, and Erik Cambria. 2024. [SarcNet: A multilingual multimodal sarcasm detection dataset](#). In *Proceedings of the*

2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024), pages 14325–14335, Torino, Italia. ELRA and ICCL.

Xiang Yue, Tianyu Zheng, Yuansheng Ni, Yubo Wang, Kai Zhang, Shengbang Tong, Yuxuan Sun, Botao Yu, Ge Zhang, Huan Sun, Yu Su, Wenhui Chen, and Graham Neubig. 2025. [MMMU-pro: A more robust multi-discipline multimodal understanding benchmark](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15134–15186, Vienna, Austria. Association for Computational Linguistics.

Rowan Zellers, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. [From recognition to cognition: Visual commonsense reasoning](#). In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6720–6731.

7 Appendix

7.1 Ablation Study: Evaluating the Performance of SOTA LLMs for Meme Classification

A natural question one may wonder is whether a dedicated meme-classification model is still necessary in the 2026- given the rapid progress of large language models (LLMs) and their increasingly strong multimodal reasoning capabilities. To examine this, we perform an evaluation using state-of-the-art multimodal LLMs for harmful meme classification on the HarMeme’s testset comprising of 354 images

For this evaluation, we select representative high-performing models from the MMMU-Pro (Yue et al., 2025) benchmark leaderboard, as visible in April 2026. From each model family, we choose the variant with the strongest reported performance. The evaluated models are: anthropic/claude-opus-4.6, openai/gpt-5.5, google/gemini-3.1-pro-preview, and moonshotai/kimi-k2.6 (Anthropic, 2026; OpenAI, 2026b; Google, 2026; Moonshot AI, 2026). Following the MMMU-Pro evaluation setting, all models are evaluated under the high-reasoning configuration, except openai/gpt-5.5, for which the medium-reasoning setting is used because it achieves superior performance under the corresponding MMMU-Pro benchmark configuration.

For each meme, we provide both the image and the associated text to the model. To ensure consistency across evaluations, we use a fixed random seed of 42. The below provided prompt is used-

DO NOT OUTPUT ANYTHING ELSE BESIDES 0 OR 1. OUTPUT ONLY FINAL ANSWER. Analyze the provided content. Determine if it qualifies as harmful according to the HarMeme definition, i.e., content that causes harm or has the potential to cause harm, including content potentially related to COVID-19 and targeting individuals, groups, or society. Output only ‘1’ if harmful and ‘0’ if not harmful based on the instructions.

Table 7 presents the performance of the evaluated LLMs on the HarMeme test set. Among the evaluated models, anthropic/claude-opus-4.6 achieves the strongest performance, with an accuracy of 0.7881 and an AUROC of 0.7775. However, the remaining models perform substantially lower, with openai/gpt-5.5, google/gemini-3.1-pro-preview, and moonshotai/kimi-k2.6 achieving accuracies of 0.6949, 0.6836, and 0.6176, respectively. The AUROC scores further indicate that several models struggle to reliably separate harmful from non-harmful memes, with moonshotai/kimi-k2.6 performing only marginally above random discrimination.

These results suggest that even highly capable general-purpose multimodal LLMs do not fully solve harmful meme classification in a zero-shot setting. Meme classification remains challenging because it often requires understanding subtle interactions between visual content, overlaid text, cultural context, sarcasm, implicit targets, and potentially harmful intent. Therefore, relying solely on off-the-shelf LLMs may be insufficient for scalable harmful meme detection.

Another important limitation is computational cost. Even on a relatively small test set of only 354 memes, the models consume a large number of tokens. The total token usage ranges from 95,842 tokens for anthropic/claude-opus-4.6 to 489,911 tokens for google/gemini-3.1-pro-preview. Such token consumption can lead to substantial inference costs when scaled to large meme datasets or real-world moderation pipelines involving thousands or millions of samples. In contrast, a dedicated task-specific model can provide a more efficient and cost-effective alternative while also being optimized directly for the harmful meme classification task.

Model	Accuracy	AUROC	Prompt Tokens	Completion Tokens	Total Tokens
anthropic/claude-opus-4.6	0.7881	0.7775	83,819	12,023	95,842
openai/gpt-5.5	0.6949	0.6073	69,461	40,732	110,193
google/gemini-3.1-pro-preview	0.6836	0.5837	344,968	144,943	489,911
moonshotai/kimi-k2.6	0.6176	0.5201	79,462	298,989	378,451

Table 7: Zero-shot harmful meme classification performance of state-of-the-art multimodal LLMs on the HarMeme test set.

7.2 Implementation Details

For all baseline models, we used the official released codebases or model checkpoints wherever available, and adapted only the dataset loading and classification heads required for our meme classification setting. Each model was trained on the corresponding training split, with the validation split used for hyperparameter tuning and best-checkpoint selection. Final performance was reported only on the held-out test split. To ensure fairness and reproducibility, experiments were run with three randomly chosen seeds. We report mean and standard deviation across seeds unless otherwise specified.

For non-VLM baselines, training was performed on an NVIDIA T4 GPU with 16GB VRAM. Vision-language model fine-tuning was performed on an NVIDIA A100 GPU with 40GB VRAM. For the VLM baselines, we used parameter-efficient fine-tuning wherever required to fit the models within the available hardware budget. As part of the experiments, for evaluation on PrideMM and HarMeme we finetune LLaVA-1.5-7B, Qwen2.5-VL. For evaluation on MemeCF, we evaluate Gemma 3 Vision (12B), LLaVA-1.5-7B, Qwen3-VL-8B and Phi-4 Multimodal in a zero-shot setting

For LLaVA-1.5-7B, we followed the official LLaVA v1.5 task-LoRA fine-tuning recipe and adapted the instruction data to our binary meme classification setting. Specifically, dataset-specific JSON files were constructed for each task, and checkpoints and logs were maintained separately for each dataset. The meme image and the corresponding text transcription were included in the prompt, allowing the model to condition on both visual content and meme text. The best checkpoint was selected using validation performance before evaluation on the test set.

For Qwen2.5-VL, we used a dedicated multimodal fine-tuning wrapper. Each example was formatted as a conversation-style binary classification instance consisting of the meme image, a fixed classification instruction, and one of the allowed

class labels as the target output. LoRA or QLoRA adapters were applied to the language-model attention and MLP layers, while the vision components were kept frozen by default. As with the other VLMs, validation macro-F1 was used for checkpoint selection.

During VLM evaluation, predictions were obtained by comparing the likelihood of the fixed candidate class labels rather than relying on unconstrained free-form generation. This produced deterministic label predictions and ensured that all VLMs were evaluated under the same binary classification protocol.

7.3 Dataset Construction & Statistics

We construct MemeCF dataset using a semi-automated pipeline that combines LLM-as-a-judge with human validation. Using images from HarMeme, PrideMM, HMC, MultiBully, MMSD2, GuardHarMeme (Pramanick et al., 2021a; Shah et al., 2024; Kiela et al., 2020; Maity et al., 2022; Qin et al., 2023; El-amrany et al., 2025)- we make use of GPT 5.5 model (OpenAI, 2026b) (having the highest performance on MMMU-Pro benchmark leaderboard at the time of dataset creation) for each meme to generate a task-specific binary label and structured counterfactual supervision. This supervision includes the modality of the decisive cue (*text*, *image*, or *both*), the minimal evidence supporting the prediction, a minimal intervention expected to flip the label, and a short rationale explaining why the proposed intervention should change the class. To reduce annotation noise we follow a strict secondary validation using GPT 5.4, Claude Opus 4.6 And Gemini Flash 3 model (OpenAI, 2026a; Anthropic, 2026; Google, 2025) as the judges- retaining only outputs with valid binary labels and confidence scores of at least 0.9. The retained pool is then manually reviewed to remove noisy, ambiguous, low-quality, or otherwise unreliable examples. Finally, accepted examples are assigned to train, validation, and test partitions

The dataset contains three task-specific subsets:

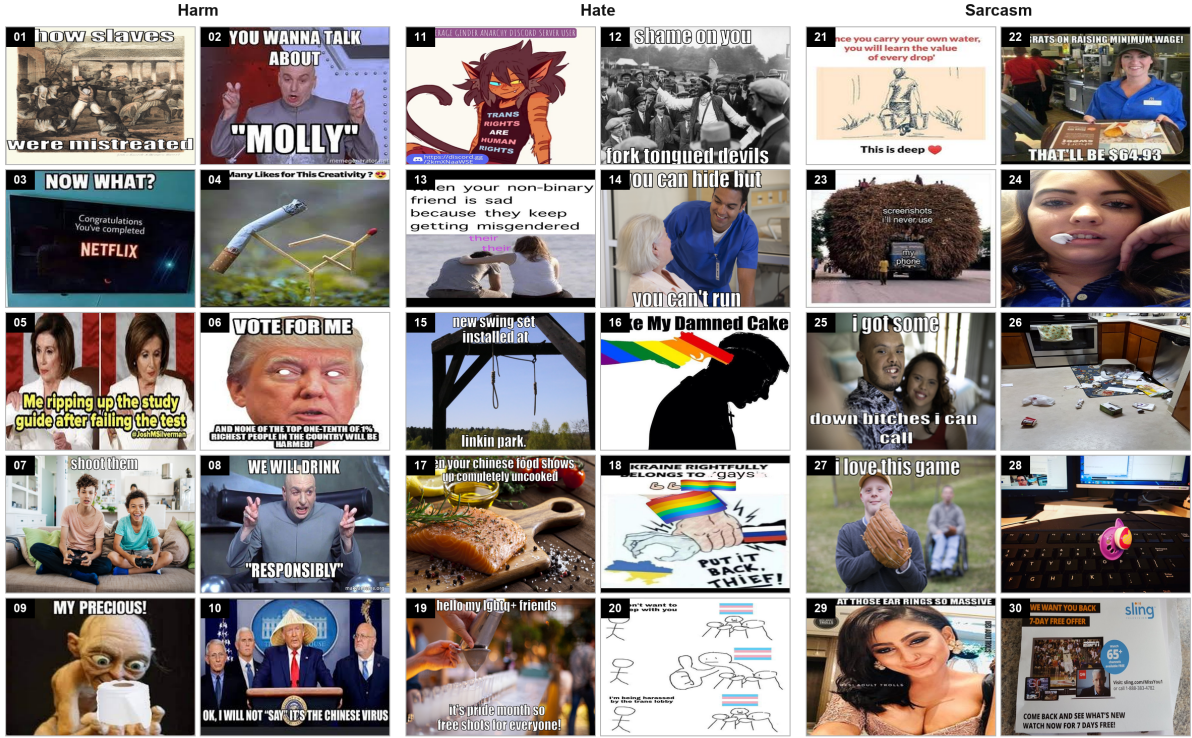


Figure 4: Thirty representative non-trivial MemeCF examples across harm, hate, and sarcasm. Within each task pair, the label-0 example appears on the left and the label-1 example on the right; the numbered examples correspond to the annotations in Table 8.

harm, hate, and sarcasm. The harm subset is derived from GuardHarMem and HarMeme, the hate subset from PrideMM and HMC, and the sarcasm subset from MMSD2 and MultiBully. Across the three tasks, the benchmark contains 9,895 memes in total, comprising 7,421 training memes, 990 validation memes, and 1,484 test memes, corresponding to an approximately 75/10/15 split.

The dataset annotations contain the meme image, associated OCR or caption text, task label, data split, source information, and counterfactual supervision. Label semantics are task-specific: for harm, 0 denotes non-harmful and 1 denotes harmful; for hate, 0 denotes non-hateful and 1 denotes hateful; and for sarcasm, 0 denotes non-sarcastic and 1 denotes sarcastic. Counterfactual supervision specifies the decisive cue, a proposed intervention that would flip the label, and a rationale explaining the expected change. The harm subset contains 3,018 memes, with 1,962 from GuardHarMem (65.0%) and 1,056 from HarMeme (35.0%). Its label-0/label-1 distribution is 1,397/1,621. The subset is partitioned into 2,263 training, 302 validation, and 453 test memes. Its counterfactual cues comprise 2,374 text-based, 411 image-based,

and 233 multimodal examples. The mean text length is 74.0 characters, with a median of 62 characters.

The hate subset contains 3,289 memes, sourced from 1,735 PrideMM memes (52.8%) and 1,554 HMC memes (47.2%). Its label-0/label-1 distribution is 1,500/1,789. The subset contains 2,467 training, 329 validation, and 493 test memes. Its counterfactual cues comprise 2,783 text-based, 377 image-based, and 129 multimodal examples. The mean text length is 108.0 characters, with a median of 73 characters.

The sarcasm subset contains 3,588 memes, including 2,029 from MMSD2 (56.5%) and 1,559 from MultiBully (43.5%). Its label-0/label-1 distribution is 1,639/1,949. The subset is divided into 2,691 training, 359 validation, and 538 test memes. Its counterfactual cues comprise 2,338 text-based, 1,100 image-based, and 150 multimodal examples. The mean text length is 58.5 characters, with a median of 53 characters. MemeCF thereby provides a standardized format and supervision style across all three tasks while preserving source diversity and task-specific label semantics. Across the full

collection, text-based counterfactuals are the most common, accounting for 7,495 memes, followed by 1,888 image-based memes and 512 multimodal memes. This design makes the benchmark suitable not only for binary meme classification, but also for counterfactual reasoning, evidence attribution, and multimodal analysis. Figure 4 shows 30 non-trivial examples and Table 8 provides the corresponding annotations.

Table 8: Thirty representative non-trivial MemeCF examples and their cue-focused annotations.

No.	Task / Type / Label	Text / Evidence	Flip / Rationale
01	Task: Harm Type: text Label: 0	Text: how slaves were mistreated Evidence: the word 'mistreated'	Flip: change text word 'mistreated' to 'disciplined' Rationale: The word 'mistreated' is the dominant local cue that makes the meme a condemnatory historical statement rather than hateful content. Replacing it with 'disciplined' reframes the same slavery scene as justified punishment, substantially changing the meme's core meaning and making it harmful toward a protected group.
02	Task: Harm Type: text Label: 1	Text: you wanna talk about "molly" Evidence: the word 'molly'	Flip: change the text word 'MOLLY' to 'MUSIC' Rationale: The quoted word 'MOLLY' is the main label-bearing cue because it is the only explicit content tied to a drug reference; the image is just a generic reaction pose. Replacing that single word with a harmless topic like 'MUSIC' keeps the meme coherent but removes the harmful drug connotation, so the label flips to non-harmful.
03	Task: Harm Type: text Label: 0	Text: NOW WHAT? Congratulations You've completed NETFLIX Evidence: the word 'NETFLIX'	Flip: change text from 'NETFLIX' to 'Jews' Rationale: The single word 'NETFLIX' is the dominant local cue that makes the meme a harmless joke about binge-watching a streaming service. Replacing that noun with 'Jews' changes the core meaning from finishing a TV catalog to 'completing' a protected group, which reads as violent/eliminationist and flips the meme to harmful.
04	Task: Harm Type: image Label: 1	Text: how many likes for this creativity? Evidence: a cigarette being kicked by a matchstick figure	Flip: Replace the cigarette on the left with a non-smoking object, such as a pencil or lollipop, while keeping the rest of the image unchanged. Rationale: The cigarette is the main local evidence for the harmful label because it directly signals smoking/tobacco use, which makes the meme harmful. Swapping it for a benign object preserves the "creative craft" idea but removes the harmful association, flipping the label to non-harmful.
05	Task: Harm Type: text Label: 0	Text: Meripping up the study guide after failing the test loshM-Silverman Evidence: study guide	Flip: change text from 'study guide' to 'Quran' Rationale: The phrase 'study guide' is the compact object being ripped up and is the main cue that the meme is a harmless school-related joke. Replacing it with 'Quran' changes the act into destruction of a religious text, shifting the meaning from frustration about a failed test to religiously hostile desecration, which would flip the label to harmful.
06	Task: Harm Type: image Label: 1	Text: VOTE FOR ME AND HONE OF THE TOP ONE-TENTH OF 1% RICHEST PEOPLE IN THE COUNTRY WILL BE HARMED! Evidence: The eyes of Donald Trump are edited to be solid white, giving a demonic or soulless appearance.	Flip: change Trump's blank white eyes to normal human eyes with visible pupils/irises Rationale: The strongest local evidence for harmfulness is the eye region: the blank white eyes make Trump look monstrous/possessed and turn the meme from ordinary political criticism into a hostile, dehumanizing depiction. Replacing just that compact visual attribute with normal eyes keeps the meme coherent but substantially changes its tone and meaning, making it a non-harmful political joke rather than a harmful attack.
07	Task: Harm Type: image Label: 0	Text: shoot them Evidence: the boys are holding video game controllers	Flip: change the pair of video game controllers in the children's hands to real guns Rationale: The dominant local evidence making the meme non-harmful is the compact visual cue that the children are playing a video game. The pair of controllers is the smallest semantically complete image region that carries most of that signal. Replacing them with real guns changes the core meaning of 'shoot them' from a likely in-game action to a real-world call for violence, which flips the label to harmful.
08	Task: Harm Type: both Label: 1	Text: we will drink "responsibly" Evidence: Text: quotation marks around the word 'responsibly'; Visual: Dr. Evil's finger air quotes gesture	Flip: remove the quotation marks around 'responsibly' and change the image's air-quote hand gesture to a neutral/non-quoting pose Rationale: The dominant harmfulness cue is the sarcastic scare-quote concept around 'responsibly', expressed both in the text quotation marks and the character's air-quote gesture. That local unit turns the statement into mockery of responsible drinking, implying irresponsible alcohol use. Replacing both cues with sincere, non-sarcastic versions changes the core meaning to genuinely drinking responsibly, which flips the meme to non-harmful.
09	Task: Harm Type: image Label: 0	Text: MY PRECIOUS! Evidence: a roll of toilet paper	Flip: change the toilet paper roll in Gollum's hands to a swastika symbol Rationale: The toilet paper roll is the smallest semantically complete visual unit that makes the meme a harmless joke about obsessing over an everyday item. Replacing that single object with a swastika substantially changes the core meaning from benign fixation to affectionate valuation of a hate symbol, which would make the meme harmful.
10	Task: Harm Type: both Label: 1	Text: OUSE OK, I WILL NOT "SAY ITS THE CHINESE VIRUS" Evidence: Text: The specific phrase 'CHINESE VIRUS' in the caption.; Visual: The edited conical straw hat (non-la) on Donald Trump's head.	Flip: change text from 'CHINESE' to 'COVID-19' and change the image conical Asian hat on Trump to his normal hair Rationale: The harmful meaning is driven by the combined China/Asian reference: the text uses the stigmatizing phrase 'Chinese virus' while the image reinforces it with a stereotyped conical Asian hat. Changing only the word or only the hat would leave the other dominant cue intact, but replacing both removes the xenophobic attribution and makes the meme a generic joke about avoiding a phrase, flipping it to non-harmful.

No.	Task / Type / Label	Text / Evidence	Flip / Rationale
11	Task: Hate Type: text Label: 0	Text: average gender anarchy discord server user trans rights are human rights 2kmxnaawse Evidence: the phrase 'trans rights are human rights'	Flip: change the shirt text from 'TRANS RIGHTS ARE HUMAN RIGHTS' to 'TRANS RIGHTS AREN'T HUMAN RIGHTS' Rationale: The shirt slogan is the dominant label-bearing cue; it directly states the meme's stance. Replacing it with an anti-trans version flips the core meaning from supportive/non-hateful to hateful while keeping the meme coherent.
12	Task: Hate Type: text Label: 1	Text: shame on you fork tongued devils Evidence: fork tongued devils	Flip: change text from 'fork tongued' to 'lying' Rationale: The two-word phrase 'fork tongued' is the smallest dominant label-bearing unit. It invokes a racialized stereotype tied to Native Americans, while the rest of the caption is mainly generic condemnation. Replacing it with 'lying' preserves a coherent insult but removes the stereotype-driven hateful meaning, which should flip the meme to non-hateful.
13	Task: Hate Type: text Label: 0	Text: when your non binary friend is sad because they keep getting misgendered their their Evidence: the word 'their' used as a pun for 'there'	Flip: change the overlaid phrase from 'their their' to 'it it' Rationale: The small overlaid phrase 'their their' is the meme's punchline and main label-bearing cue: it reframes the comforting phrase 'there there' with a non-binary pronoun, making the meme sympathetic rather than hostile. Replacing that phrase with 'it it' changes the core meaning from supportive wordplay to dehumanizing misgendering of a non-binary person, which would make the meme hateful.
14	Task: Hate Type: image Label: 1	Text: you can hide but you can't run Evidence: the presence of an elderly woman who appears to have limited mobility	Flip: change the elderly seated woman to a clearly younger adult woman in the same pose and setting Rationale: The dominant label-bearing cue is the image's depiction of an elderly target. The text is generic on its own, but paired with an obviously old woman it becomes an age-targeted taunt about her inability to run, which makes the meme hateful. Replacing the elderly woman with a younger adult removes the protected-age targeting while keeping the scene coherent, flipping the meme to non-hateful.
15	Task: Hate Type: text Label: 0	Text: new swing set installed at linkin park Evidence: linkin park	Flip: change text from 'linkin park' to 'the mosque' Rationale: The phrase 'linkin park' is the key local cue that makes the noose image read as a dark joke about the band Linkin Park/Chester Bennington rather than as hostility toward a protected group. Replacing only that phrase with 'the mosque' keeps the meme coherent but changes the target to a Muslim religious site, so the hanging image becomes an implied anti-Muslim threat and the label flips to hateful.
16	Task: Hate Type: image Label: 1	Text: bake my damned cake Evidence: the rainbow colors (LGBTQ+ flag) applied to the gun silhouette	Flip: change the rainbow pride-colored stripes hitting the silhouette's head to a plain neutral gray paint streak Rationale: The rainbow pride-colored visual cue is the dominant local evidence identifying the target as LGBTQ people and framing them as aggressively demanding service. The text alone is just an angry demand about cake; without the pride cue, the meme becomes a generic complaint rather than anti-LGBTQ. Replacing the rainbow with a neutral paint streak removes the protected-class targeting and flips the meme to non-hateful.
17	Task: Hate Type: image Label: 0	Text: when your chinese food shows up completely uncooked Evidence: a piece of raw salmon	Flip: change the raw salmon fillet on the cutting board to a clearly recognizable dead dog on the cutting board Rationale: The salmon fillet is the dominant local visual cue that makes the meme a benign joke about raw food. Replacing that single food item with a recognizable dog would substantially change the meaning from 'uncooked meal' to the racist stereotype that Chinese food is dog meat, which would make the meme hateful toward Chinese people.
18	Task: Hate Type: both Label: 1	Text: ukraine rightfully belongs to gays bb put it back thief Evidence: Text: The word 'gays' in the top caption; Visual: The rainbow flag armbands and icons overlaid on the muscular arm	Flip: change text from 'gays' to 'ukrainians' and change the image rainbow flag on the fist to a Ukrainian flag Rationale: The local evidence identifying the aggressor as gay/LGBTQ is carried by the word 'gays' together with the rainbow flag on the fist. This protected-class cue is the dominant reason the meme is hateful, because it frames gays as thieves who wrongfully take Ukraine. Replacing that paired cue with 'ukrainians' and a Ukrainian flag changes the target from a protected group to a national self-ownership message, making the meme non-hateful.
19	Task: Hate Type: image Label: 0	Text: hello my lgbtq+ friends it's pride month so free shots for everyone! Evidence: the liquid being poured from the shaker	Flip: change the local drink-shot region (the row of shot glasses being served/poured) to a handgun being aimed or fired instead of drink shots Rationale: The shot-glass serving region is the smallest dominant cue that disambiguates 'shots' as alcohol, which makes the meme a pride-month celebration and therefore non-hateful. Replacing that compact visual unit with a gun/firearm scene would reinterpret 'free shots' as gunshots directed at the mentioned LGBTQ+ group, substantially changing the meme's meaning and flipping the label to hateful.
20	Task: Hate Type: both Label: 1	Text: i don't want to sleep with you i'm being harassed by the trans lobby 11 Evidence: Text: The phrase 'trans lobby' in the bottom text.; Visual: The transgender pride flag appearing above the group of stick figures.	Flip: change text from 'trans lobby' to 'dating app lobby' and change the trans pride flags to generic dating app icons Rationale: The dominant hateful signal is the local identification of the harassing group as trans, reinforced by the trans pride flags. Replacing that protected-group cue with a non-protected generic dating-app cue changes the meme from portraying trans people as sexual harassers to a non-hateful joke about app-driven harassment, flipping the label.

No.	Task / Type / Label	Text / Evidence	Flip / Rationale
21	Task: Sarcasm Type: text Label: 0	Text: Once you carry your own water,you will learn the value of every drop'This is a deep Evidence: the heart emoji at the bottom	Flip: change the heart emoji in 'This is deep ♡' to a rolling-eyes emoji Rationale: The heart emoji is the smallest local tone cue making the bottom caption read as sincere appreciation of the inspirational quote. Replacing it with a rolling-eyes emoji turns 'This is deep' into a mocking reaction, substantially changing the meme from earnest motivational content to sarcastic dismissal.
22	Task: Sarcasm Type: text Label: 1	Text: wouldn 't it be awesome if everyone could make \$ 75 per hour ? Evidence: \$64.93	Flip: Change the text "\$64.93" to "\$6.49". Rationale: The exaggerated price is the main sarcastic punchline. Replacing it with a normal fast-food price removes the irony, making the meme read as a straightforward transaction rather than a sarcastic complaint.
23	Task: Sarcasm Type: text Label: 0	Text: Screenshots I'll never use my phone Evidence: the word 'never' in the phrase 'screenshots i'll never use'	Flip: change text from 'never' to 'always' Rationale: The word 'never' is the smallest dominant text unit that makes the meme a literal, non-sarcastic complaint about storing many useless screenshots. Replacing it with 'always' creates a clear contradiction between the enormous pile of screenshots and the claim that they will all be used, making the meme sarcastic.
24	Task: Sarcasm Type: image Label: 1	Text: man i love the dentist . Evidence: the white cotton ball/gauze tucked into the side of the woman's mouth	Flip: replace the cotton/gauze wad sticking out of her mouth with a normal closed mouth or relaxed smile Rationale: The protruding dental gauze is the dominant local cue that signals an unpleasant just-left-the-dentist situation, which directly contradicts the positive text 'man i love the dentist .' and creates the sarcasm. Removing that single visual cue makes the selfie no longer depict obvious dental discomfort, so the text reads as straightforward rather than sarcastic.
25	Task: Sarcasm Type: text Label: 0	Text: i got some down bitches i can call Evidence: the word 'down'	Flip: change the word 'down' to 'fine' Rationale: The single word 'down' is the dominant local cue because it makes the caption align with the pictured woman as a Down-syndrome joke rather than an ironic boast. That alignment is what keeps the meme non-sarcastic. Replacing 'down' with 'fine' changes the core meaning to bragging about attractive women, which the image then undercuts, creating sarcasm.
26	Task: Sarcasm Type: image Label: 1	Text: love it when my mothers dogs are here . Evidence: the knocked-over trash can and scattered garbage in the kitchen	Flip: change the overturned trash can and scattered garbage on the kitchen floor to a clean, undisturbed floor with the trash can upright Rationale: The dominant evidence for sarcasm is the local visual mess caused by the dogs: the overturned bin and spilled trash directly contradict the positive text 'love it when my mothers dogs are here.' This compact visual region carries most of the sarcastic signal. Replacing it with a clean floor and upright bin changes the implied behavior of the dogs from destructive to harmless, making the text read as sincere rather than sarcastic.
27	Task: Sarcasm Type: image Label: 0	Text: i love this game Evidence: the smiling, happy facial expression of the boy with Down syndrome	Flip: change the boy's calm/happy face to a crying or visibly upset face Rationale: The boy's positive facial expression is the main local visual cue that makes the caption 'i love this game' read sincerely. Replacing only his face with an upset expression creates a clear contradiction between the enthusiastic text and the negative reaction, making the meme read sarcastically.
28	Task: Sarcasm Type: both Label: 1	Text: ways to impress higher ups : reach for a pen and pull a paci out of your pocket . Evidence: Text: the word 'paci' in the text; Visual: the purple pacifier sitting on the keyboard	Flip: change text from 'paci' to 'pen' and change the image object from a pacifier to a pen Rationale: The pacifier is the dominant local cue carrying the sarcasm: it makes the supposed advice for impressing higher-ups obviously childish and counterproductive. Replacing that single entity with a pen changes the core meaning from embarrassing yourself to appearing prepared/professional, which makes the meme read as non-sarcastic advice.
29	Task: Sarcasm Type: image Label: 0	Text: look at those ear rings so massive Evidence: the size of the earrings on the woman's ears	Flip: change the woman's large flower-shaped earrings to tiny plain stud earrings Rationale: The dominant evidence for the current non-sarcastic label is the visible oversized earrings, which make the caption's literal claim coherent. The earrings are a compact, semantically complete visual unit. Replacing them with tiny studs would make the text 'look at those ear rings so massive' clearly contradict the image, substantially changing the meme's meaning and flipping it to sarcastic.
30	Task: Sarcasm Type: both Label: 1	Text: oh wow ! thank u so much <user> you must really want me back . <num> whole days ? ! what an offer ! icymi Evidence: Text: The phrase '7 whole days'; Visual: The text '7-DAY' on the physical mailer	Flip: change text from '<num> whole days' to '90 whole days' and change the image offer duration from '7-DAY/7 DAYS FREE' to '90-DAY/90 DAYS FREE' Rationale: The dominant local evidence is the short 7-day free-offer duration, which appears in both the meme text and the flyer. The sarcasm comes from treating a very small win-back offer as if it were generous. Replacing the duration with a much more generous 90-day offer makes the praise coherent and sincere, flipping the meme to non-sarcastic.