MULTI-QUADRANT SYNTHETIC DATA-DRIVEN CONTRASTIVE LEARNING FOR HATEFUL MEME DETECTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Internet memes express users' viewpoints through metaphors and may carry implicit hateful messages, making hateful meme detection challenging. Mainstream methods embed image and text features into a joint space and classify a meme as hateful when its pattern matches known hateful features, such as a strong correlation between malicious text and inflammatory images. However, they ignore a key property of memes: two memes with highly similar surface content may convey completely opposite metaphors. Such confounding memes lie close to each other in the feature space but carry opposite labels, so the two classes mix and become hard to distinguish. We propose a Multi-Quadrant Synthetic data-driven Contrastive Learning (MQSCL) framework that characterizes any meme pair along two content-level dimensions, surface similarity and metaphor-polarity consistency, and uses metaphor polarity to supervise the decision boundary. It consists of three modules: (1) Metaphor Rationale Analysis: a multi-agent pipeline combines multimodal large language model interpretation with background knowledge retrieval to build a metaphor reasoning graph, locate the targeted group, and generate a metaphor rationale; (2) Quadrant Counterpart Generation: conditioned on the rationale, a diffusion model with a text-image decoupled strategy generates, for each source meme, four quadrant counterparts covering all combinations of the two dimensions; (3) Multi-Quadrant Contrastive Learning: a Quadrant Calibration Loss (QCL) with regression, repulsion, and ordering terms calibrates the anchor-counterpart similarities, so that the decision boundary is determined by metaphor polarity rather than surface similarity. The counterparts are used only during training and introduce no additional inference cost. On standard benchmarks, MQSCL outperforms the strongest detectors; on the most challenging FHM dataset, its F1 is 3.9 points higher than the state of the art. Caution: Contains academic discussions of hate speech; viewer discretion advised.

1 INTRODUCTION

Internet memes combine images with short text and rely on metaphor to convey specific intents, mostly in a humorous form. However, when the conveyed message turns hateful (reinforcing stereotypes about certain groups, dehumanizing them, or inciting violence), social media platforms must detect and address it in a timely manner. Hateful meme detection is therefore an important problem in multimodal content understanding, attracting broad attention from both academia and industry Kiela et al. (2020); Pramanick et al. (2021a); Kumar & Nandakumar (2022); Lu et al. (2025); Cao et al. (2022).

Existing methods mainly model image-text interactions: a meme is classified as hateful when its pattern matches known hateful features, such as a strong correlation between malicious text and inflammatory images Kiela et al. (2020; 2019); Pramanick et al. (2021b). These methods face a key obstacle: *confounding memes*, pairs of memes whose surface content is almost identical but whose metaphors are completely opposite. This phenomenon is not an anecdote: it was first identified and empirically confirmed by RGCL Mei et al. (2024), whose analysis shows that such opposite-labeled near-duplicates are widespread in standard benchmarks and are a consistent source of misclassification for interaction-based detectors. As shown in Figure 1(a), the two memes share a similar visual

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

category and the same text semantics, yet convey completely opposite metaphors. The root cause lies in how detectors encode a meme as a joint image-text embedding: confounding memes lie close to each other in the feature space yet carry opposite labels, compressing the decision boundary into a very narrow region where similarity-based discrimination fails.

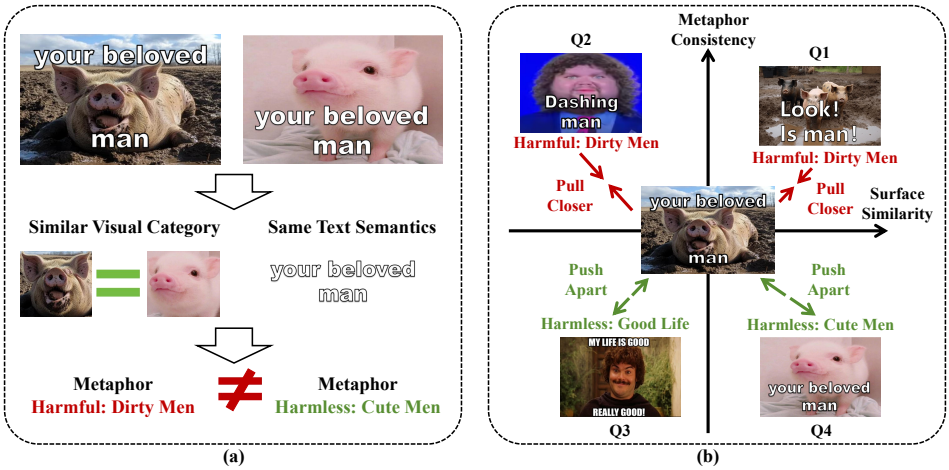

Figure 1: The confounding-meme problem and the quadrant framework. (a) A confounding pair: nearly identical image and text, yet opposite metaphors. (b) Meme pairs fall into four quadrants along surface similarity and metaphor-polarity consistency; Q1 and Q3 are easy for similarity-based discrimination, while Q2 and Q4 are where it fails. MQSCL synthesizes one counterpart per quadrant for each anchor and reorganizes the space so that feature distances follow metaphor polarity rather than surface similarity.

To analyze this problem systematically, we characterize any meme pair along two content-level axes: *surface similarity* (whether the two memes share similar visual content and text semantics) and *metaphor-polarity consistency* (whether the pair carries the same metaphor polarity, i.e., both hateful or both non-hateful). Being content-level, the axes are independent of any particular embedding space; under a frozen encoder, surface-similar memes appear as neighboring points, which is why confounding memes cluster near the decision boundary. The two axes induce four quadrants (Figure 1(b)): (Q1) similar surface, consistent polarity: an easy positive pair; (Q2) dissimilar surface, consistent polarity: same label but far apart in the feature space, so metric-based supervision fails to bind it together; (Q3) dissimilar surface, opposite polarity: an easy negative pair; (Q4) similar surface, opposite polarity: the confounding case, close in the feature space yet on opposite sides of the boundary, leaving only a tiny margin. Q1 and Q3 are handled naturally, whereas Q2 and Q4 are the true sources of confusion: in Q4, surface similarity acts as a shortcut that misleads the decision boundary; in Q2, semantic identity cannot be inferred from feature-space distance. An ideal feature space is therefore organized by metaphor polarity rather than surface similarity: Q1 stays closest to the anchor, Q2 is pulled close despite its distant features, Q4 is pushed away despite its close features, and Q3 is pushed the farthest.

We propose Multi-Quadrant Synthetic data-driven Contrastive Learning (MQSCL). Its core idea is to locate the misalignment between surface similarity and metaphor-polarity consistency (Q2 and Q4) in the feature space and resolve it with synthetic data: for each source meme, four quadrant counterparts are synthesized, and the feature space is reorganized with metaphor polarity as the geometric supervision, so that the decision boundary follows metaphor polarity rather than surface similarity. The framework consists of three modules (Figure 2):

(1) **Metaphor Rationale Analysis**. Three cooperating agents run sequentially on the same multimodal large language model (MLLM) backbone: a *content analysis and knowledge search agent* interprets the image, the text, and the metaphor label, and retrieves sociocultural knowledge from the web; a *graph construction agent* organizes the retrieved knowledge into an explicit, queryable metaphor reasoning graph; and a *rationale generation agent* localizes the target group along relation paths and distills the metaphor rationale. The module runs only during training-data construction

and conditions each stage on the ground-truth label, keeping the rationale consistent with the meme's polarity.

(2) **Quadrant Counterpart Generation**. Manual construction requires deep social-context understanding and is prohibitively expensive, so we generate instead of annotate: a *quadrant prompt agent* parses the source image, text, and rationale into a per-quadrant specification (image prompt, overlaid text, overlay position); a diffusion model then synthesizes the text-free background under a text-image decoupled strategy, and the text is rendered onto it, yielding one counterpart per quadrant with explicit control along both surface similarity and metaphor polarity.

(3) **Multi-Quadrant Contrastive Learning**. The Quadrant Calibration Loss (QCL) adds regression, repulsion, and ordering terms on top of the anchor's classification loss: the regression term calibrates polarity-consistent counterparts toward preset target similarities; the repulsion term pushes polarity-opposite counterparts below upper similarity bounds, driving surface-similar confounding pairs out of the anchor's neighborhood; and the ordering term maintains a near-to-far order across surface similarities within each polarity region. The three terms arrange samples by metaphor polarity, and the quadrant counterparts act only as training-time scaffolding with no additional inference cost.

Our main contributions are summarized as follows: (1) We reduce the confounding-meme problem to a four-quadrant structure along surface similarity and metaphor-polarity consistency, and accordingly propose a synthetic-data solution that automatically builds four quadrant counterparts for each source meme, avoiding expensive manual annotation. (2) We design the Quadrant Calibration Loss (QCL), a multi-quadrant contrastive objective composed of regression, repulsion, and ordering terms that calibrates the feature space according to the quadrant structure, so that the decision boundary depends on metaphor polarity rather than surface similarity. (3) Experiments on standard benchmarks show that MQSCL outperforms the strongest existing detectors; on the most challenging FHM dataset, it improves F1 over the state of the art by 3.9 points.

2 Related Work

Research on multimodal interaction primarily follows two paradigms: dual-stream architectures that classify based on visual and textual feature fusion Kiela et al. (2020; 2019); Pramanick et al. (2021b); Cao et al. (2022); Su et al. (2025), and domain adaptation of pre-trained multimodal models Lippe et al. (2020); Pramanick et al. (2021a); Yang et al. (2022); Cao et al. (2022); Mei et al. (2024); Muennighoff (2020); Sharma et al. (2022). Despite strong performance, their decision-making processes suffer from insufficient interpretability; to address this, one study Hee et al. (2022) proposes a reverse reasoning framework revealing how image-text associations form in hate content recognition. Data-centric attempts also face obstacles: datasets with metaphorical-content annotations Hee et al. (2023) remain small because manual annotation is costly, limiting model generalization.

Recent research leverages pre-trained vision-language models or large language models to generate auxiliary knowledge that enhances hate content detection and generalization Cao et al. (2023); Lin et al. (2023; 2024); Hee & Lee (2025); Lu et al. (2025), but often treats the large model as a single expert, so the generated reasoning knowledge is homogeneous and lacks diversity.

3 Method

3.1 Metaphor Rationale Analysis

Three agents (a content analysis and knowledge search agent, a graph construction agent, and a rationale generation agent) run sequentially on the same multimodal large language model (MLLM) backbone. Given a meme $m = (v, t, y)$ with image v, embedded text t, and metaphor label $y \in \{\text{hateful}, \text{non-hateful}\}$ (the polarity of the meme's metaphor), the module proceeds through content interpretation with knowledge retrieval, metaphor reasoning graph construction, and target localization with rationale generation. It runs only during training-data construction, conditioning each stage on the ground-truth label y so that the resulting understanding supports subsequent counterpart synthesis.

Content interpretation and knowledge retrieval. Metaphor understanding depends heavily on external background knowledge. Following prompt p_1 (full text in Appendix D), the content analysis and knowledge search agent produces, from the image, the text, and the metaphor label, a content

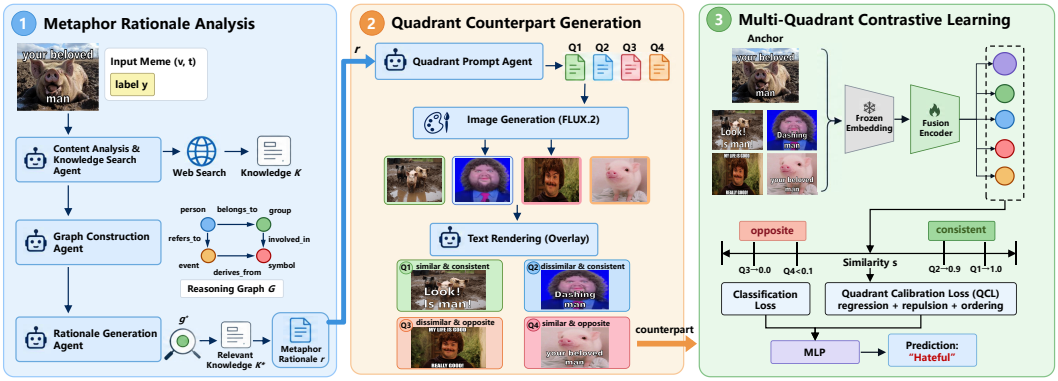

Figure 2: The overall framework of MQSCL: (1) Metaphor Rationale Analysis: three MLLM agents retrieve sociocultural knowledge, organize it into a metaphor reasoning graph, and generate a metaphor rationale; (2) Quadrant Counterpart Generation: the rationale is converted into per-quadrant specifications, from which counterparts are synthesized through text-free diffusion and text rendering; (3) Multi-Quadrant Contrastive Learning: the Quadrant Calibration Loss (QCL) calibrates the anchor-counterpart similarities so that the decision boundary follows metaphor polarity.

interpretation together with the background knowledge retrieved in the same call:

$$(c, \mathcal{K}) = \mathrm{MLLM}(v, t, y; p_1) \tag{1}$$

where c is the content interpretation—the list of related elements output by the agent, covering the main visual subjects, key textual entities, and underlying sociocultural symbols—and $\mathcal{K} = \{k_1, k_2, \ldots, k_n\}$ collects the retrieved background-knowledge summaries about people, groups, events, and cultural symbols, each with a traceable source; the web search is grounded in this interpretation through the search tool attached to the agent.

Metaphor reasoning graph construction. The retrieved knowledge is unstructured and cannot directly support multi-hop reasoning, so the graph construction agent organizes $\mathcal{K}$ into a metaphor reasoning graph following prompt p_2 (see Appendix D):

$$\mathcal{G} = (\mathcal{V}, \mathcal{E}) = \mathrm{MLLM}(\mathcal{K}; p_2), \tag{2}$$

where the node set $\mathcal{V}$ covers four entity types (person, group, event, symbol) with group nodes forming $\mathcal{V}_{\text{group}}$, and the edge set $\mathcal{E} \subseteq \mathcal{V} \times \mathcal{R} \times \mathcal{V}$ encodes relations $\mathcal{R} = \{\text{belongs_to}, \text{refers_to}, \text{involved_in}, \text{derives_from}\}$. The graph thereby turns the meme's implicit social context into structured knowledge that supports reasoning.

Target localization and rationale generation. The rationale generation agent first localizes the target group in the graph following prompt p_3:

$$g^* = \mathrm{Locate}\big(v, t, y, \mathcal{G}; p_3\big) \subseteq \mathcal{V}_{\text{group}}, \tag{3}$$

where the rationale generation agent localizes the target group purely by reasoning over the evidence-backed graph under prompt p_3: it traces relation paths from person, symbol, and event nodes to group nodes and returns the group or groups most closely associated with the meme's persons, symbols, and events—no external numeric scorer is involved. No group is assigned ($g^* = \varnothing$) when the graph contains no group node or no candidate group is supported by the graph evidence, which usually occurs on non-hateful memes that involve no social group. The agent then selects, from the retrieved knowledge, the items relevant to g^* along relation paths (see Appendix D):

$$\mathcal{K}^* = \mathrm{Select}\big(\mathcal{K},\, g^*,\, \mathcal{G}; p_3\big) \tag{4}$$

where the agent inspects the relation paths around g^* in $\mathcal{G}$ and keeps the knowledge items whose entities connect to the localized group, again by its own reasoning rather than a numeric score; when $g^* = \varnothing$, it keeps the knowledge directly related to the meme content itself. Finally, the agent generates the rationale under the joint constraint of the meme and the relevant knowledge:

$$r = \mathrm{MLLM}(v, t, y, g^*, \mathcal{K}^*; p_3) \tag{5}$$

where r explains how the meme invokes, maps to, or depicts the target group through visual elements and textual cues, focusing on negative stereotypes, derogatory expressions, or discriminatory metaphor mechanisms for hateful memes, and on humorous or ordinary semantic content for non-hateful ones (including those with $g^* = \varnothing$).

3.2 Quadrant Counterpart Generation

Based on the metaphor rationale r, this module constructs four quadrant counterparts for each source meme through three steps: quadrant prompt generation, text-free image synthesis, and text rendering.

Quadrant prompt generation. The four quadrants are defined by two control dimensions: surface similarity to the source meme (similar or dissimilar) and metaphor-polarity consistency (consistent or opposite), giving Q_1 (similar and consistent), Q_2 (dissimilar and consistent), Q_3 (dissimilar and opposite), and Q_4 (similar and opposite). The quadrant prompt agent, likewise an MLLM, takes the source image v, text t, and rationale r, parses the visual style, subjects, scene structure, and text layout under prompt p_4 (full text in Appendix D), and produces a structured generation specification for each target quadrant q:

$$(P_q, \hat{t}_q, b_q) = \mathrm{MLLM}(v, t, r, q; p_4), \quad q \in \{Q_1, Q_2, Q_3, Q_4\}, \tag{6}$$

where P_q is a concise natural-language image prompt, $\hat{t}_q$ the overlaid text, and b_q its overlay position. For polarity-consistent quadrants (Q_1, Q_2), the specification preserves the referents and metaphor polarity described by r and only keeps or changes the surface content; for polarity-opposite quadrants (Q_3, Q_4), it flips the meme's metaphor polarity so that the counterpart expresses the opposite metaphor under a similar or dissimilar surface.

Text-free image synthesis. Diffusion models often produce garbled characters when rendering text directly, so we adopt a text-image decoupled strategy: the FLUX.2 Klein-4B Labs (2025) text-to-image model $\mathcal{D}$ generates only text-free background images, guided by a negative prompt that suppresses text, captions, and lettering—visual subjects such as people are not affected—so that the model focuses on visual content and scene structure:

$$\hat{v}_q = \mathcal{D}(P_q, P_{\mathrm{neg}}), \tag{7}$$

where P_{neg} is the text-suppressing negative prompt and the output resolution is 768×768.

Text rendering. The overlaid text is then rendered onto the background image at the overlay position, yielding the quadrant counterpart:

$$\hat{m}_q = \mathrm{Overlay}(\hat{v}_q, \hat{t}_q, b_q). \tag{8}$$

3.3 Multi-Quadrant Contrastive Learning

Multi-quadrant contrastive learning injects the quadrant structure into the feature space: metaphor polarity governs the global anchor-counterpart separation, while surface similarity only determines the ordering within each polarity region. The module proceeds through multimodal feature encoding, quadrant similarity calibration, and objective optimization, and outputs a hateful or non-hateful prediction at inference.

Multimodal feature encoding. The metaphor of a meme lives in the image-text interaction, so a frozen multimodal embedding model Φ embeds the two modalities into a joint feature space and a fusion encoder $\mathcal{T}$ aggregates them:

$$(f, z) = \mathcal{T}\big(\Phi(v), \Phi(t)\big), \tag{9}$$

where z denotes the classification logits over $\{\text{hateful}, \text{non-hateful}\}$. Inside $\mathcal{T}$, the two modality embeddings are layer-normalized, augmented with modality-specific markers, and arranged with a learnable `[CLS]` token into a three-token sequence processed by a pre-norm Transformer Vaswani et al. (2017); the ℓ_2-normalized `[CLS]` output yields the fused representation f, and a two-layer MLP head produces z. We instantiate Φ as GME Zhang et al. (2024), a general multimodal embedding model (2B version, 1536-dimensional embeddings), and $\mathcal{T}$ as a 4-layer pre-norm Transformer (hidden size 1536, 16 heads, FFN width 1536 with GELU activation, dropout 0.3). Full architecture and optimization hyperparameters are listed in Appendix C.

Quadrant similarity calibration. Encoding each training anchor m (the source meme) and its four quadrant counterparts with Eq. equation 9 (a counterpart takes $\hat{v}_q$ as image and $\hat{t}_q$ as text) yields the anchor feature f and counterpart features $\hat{f}_q$, with quadrant similarity measured by the inner product

$$s_q = f^\top \hat{f}_q, \quad q \in \{Q_1, Q_2, Q_3, Q_4\}, \tag{10}$$

which equals the cosine similarity since both features are ℓ_2-normalized. To make the decision boundary depend on metaphor polarity rather than surface similarity, we design the *Quadrant Calibration Loss* (QCL), a composite calibration loss with regression, repulsion, and ordering terms:

$$\begin{aligned} \mathcal{L}_{\text{reg}} &= 2\,\mathbb{E}\big[\,|s_{Q_1} - 1| + |s_{Q_2} - 0.9|\,\big], \\ \mathcal{L}_{\text{rep}} &= \mathbb{E}\big[\,[s_{Q_4} - 0.1]_+ + [s_{Q_3}]_+\,\big], \\ \mathcal{L}_{\text{ord}} &= \mathbb{E}\big[\,[\,0.1 - (s_{Q_1} - s_{Q_2})\,]_+ \\ &\qquad + [\,0.05 - (s_{Q_4} - s_{Q_3})\,]_+\,\big], \end{aligned} \tag{11}$$

where $[x]_+ = \max(x, 0)$ and the expectation is over training memes. The regression term calibrates consistent counterparts to the targets (1.0 and 0.9); the two-sided ℓ_1 gives a constant-magnitude gradient on either side of the target and prevents overshooting it and collapsing the similar-dissimilar gap, and the factor 2 up-weights this term relative to the repulsion and ordering terms, reflecting that calibrating counterparts onto exact targets is harder than bounding them from above. The repulsion term pushes opposite counterparts below the upper bounds (0.1 and 0.0), expelling surface-similar confusers from the anchor neighborhood; its one-sided hinge stops pushing once inactive, as pushing farther brings no additional gain and would only over-compress the feature space. The ordering term preserves the within-region surface ordering ($s_{Q_1} > s_{Q_2}$, $s_{Q_4} > s_{Q_3}$) even before the targets are reached. The wide consistent-opposite gap thus encodes the metaphor polarity, the within-region ordering encodes surface similarity, and the decision boundary follows metaphor polarity rather than surface similarity.

Overall objective and inference. QCL combines with the classification supervision on the anchor:

$$\mathcal{L} = \mathcal{L}_{\text{cls}} + \lambda\,\mathcal{L}_{\text{QCL}}, \qquad \mathcal{L}_{\text{QCL}} = \mathcal{L}_{\text{reg}} + \mathcal{L}_{\text{rep}} + \mathcal{L}_{\text{ord}}, \tag{12}$$

where $\mathcal{L}_{\text{cls}}$ is the cross-entropy on the anchor logits and λ is empirically set to 0.5, balancing the typically larger magnitude of the QCL term. During training, the anchor and its four quadrant counterparts share one batch: the classification loss supervises only the anchor, while QCL involves all counterparts through their similarities to the anchor. At inference only the input meme is encoded and classified, so the quadrant counterparts are training-time scaffolding with no inference overhead.

4 Experiments

We answer six research questions: **RQ1**: How does MQSCL compare with state-of-the-art detectors on FHM, Harm-C, and Harm-P? **RQ2**: Are the generated metaphor rationales faithful to human annotations? **RQ3**: Do the quadrant counterparts follow the quadrant specification? **RQ4**: How much do the synthetic counterparts and QCL each contribute? **RQ5**: Are the improvements stable across seeds and statistically significant? **RQ6**: How do the rationale, counterparts, and learned geometry behave on confounding memes?

4.1 Experimental Settings

Implementation Details: MQSCL follows the models and training configuration of Sections 3.1–3.3; the MLLM behind the three rationale-analysis agents and the quadrant prompt agent is instantiated as Gemini-3-Flash Team et al. (2025). Hyperparameters are detailed in Appendix C. For controlled comparison, all re-implemented baselines (the CE-only baseline of Table 6 and the RGCL re-runs of Table 7) likewise extract features with the same frozen GME-2B encoder and fine-tune a standard Transformer fusion head of the same size as ours, so every comparison in the ablation and significance analysis is decoupled from encoder capacity and model scale.

Datasets: We evaluate on FHM Kiela et al. (2020) (10K memes; standard 8,500/500/1,000 train/validation/test split) and Harm-C Pramanick et al. (2021a) and Harm-P Pramanick et al.

Table 1: Overall performance (ACC and F1, %) on FHM, Harm-C, and Harm-P

			FHM		Harm-C		Harm-P	
Type	**Model**	**Year**	**ACC**	**F1**	**ACC**	**F1**	**ACC**	**F1**
Unimodal	Text BERT Devlin et al. (2019)	NAACL 2019	57.1	41.5	70.2	66.3	80.1	78.4
Unimodal	Image-Region He et al. (2016)	CVPR 2016	52.3	34.2	68.7	63.0	73.1	72.8
Multimodal	MOMENTA Pramanick et al. (2021b)	EMNLP 2021	61.3	57.5	83.8	82.8	86.8	85.3
Multimodal	PromptHate Cao et al. (2022)	EMNLP 2022	72.9	65.2	84.5	81.5	87.1	84.1
Multimodal	HateSieve Su et al. (2025)	NAACL 2025	73.5	71.6	83.6	83.1	88.3	88.0
Knowledge-enhanced	Pro-Cap Cao et al. (2023)	ACM MM 2023	74.9	71.7	85.0	83.2	87.7	86.9
Knowledge-enhanced	IntMeme Hee & Lee (2025)	AAAI 2025	71.5	71.2	81.9	80.4	87.3	86.1
Knowledge-enhanced	ExplainHM Lin et al. (2024)	WWW 2024	74.3	74.0	85.2	83.7	88.0	86.0
Knowledge-enhanced	RGCL Mei et al. (2024)	ACL 2024	78.8	77.3	86.1	85.5	87.6	87.1
Multi-agent reasoning	M2KE Lu et al. (2025)	SIGIR 2025	75.7	75.6	86.3	85.0	88.1	86.5
Multi-agent reasoning	CapSRA Yang et al. (2026)	PR 2026	75.8	72.8	86.1	81.9	-	-
Multi-agent reasoning	ALARM Lang et al. (2026)	KDD 2026	75.8	75.8	-	-	-	-
MLLM	Gemini-3-Flash Team et al. (2025)	2025	76.3	76.0	64.4	50.0	62.8	58.2
MLLM	GPT-5.5 Barman et al. (2026)	2026	75.9	75.5	61.5	50.1	58.6	52.8
MLLM	Qwen-3.5-Plus Team (2026)	2026	76.1	75.9	63.2	44.9	55.4	45.3
Synthetic data	MQSCL (Ours)	2026	**81.6**	**81.2**	**89.6**	**87.1**	**91.4**	**90.7**

(2021b) (memes about celebrities and politicians whose harmfulness hinges on entity-specific sociopolitical context). Quadrant counterparts are generated for all training memes via the pipeline in Section 3.2. Dataset statistics are summarized in Appendix A.

Evaluation Metric: Following previous works Kiela et al. (2020); Mei et al. (2024), we report accuracy (ACC) and the F1 of the hateful class in percent.

4.2 Overall Performance (RQ1)

As shown in Table 1, MQSCL achieves the best ACC and F1 on all three benchmarks: 81.6/81.2 ACC/F1 on FHM (+2.8/+3.9 over the strongest baseline RGCL) and 89.6/91.4 ACC on Harm-C/Harm-P (+3.3/+3.1 over the best published results). Two observations stand out. First, zero-shot MLLMs are competitive on FHM ($\approx$76 ACC) yet collapse on the Harm splits (F1 near chance): FHM metaphors rest on widely known stereotypes recoverable from parametric knowledge, whereas the Harm splits hinge on entity-specific context, echoing our motivation that metaphor understanding requires grounded knowledge. Second, MQSCL surpasses even the context-injecting families (knowledge-enhanced and multi-agent detectors) with only a frozen encoder and a lightweight fusion head; since our re-implemented baselines share the same frozen encoder and a same-sized head (Section 4), the operative factor is the quadrant geometry rather than model capacity.

4.3 Quality of Metaphor Rationales (RQ2)

A metaphor rationale explains the *invisible* mechanism (which group is targeted, which stereotype is invoked, and why the image-text combination becomes hateful), so many formulations can be valid; we therefore measure semantic agreement on HatReD Hee et al. (2023), which provides human-annotated metaphor reasons for 3,228 memes in FHM. Table 2 reports BERTScore Zhang et al. (2019) against human references: adding the metaphor label ($+L$), the retrieved sociocultural knowledge ($+S$), and graph-structured reasoning ($+G$) to the MLLM baseline lifts BERT-F from 0.239 to 0.384, closing 80% of the gap to the supervised VL-T5$_{\mathrm{HatReD}}$ (0.420). As BERTScore favors outputs that imitate the reference style, we complement it with 5-point Likert Hee et al. (2023) ratings from five third-party MLLMs and five human experts on a fixed-seed random sample of 200 rationale–reference pairs from HatReD (Table 3; protocol in Appendix B), where the full module is rated highest by both panels (4.88/4.04), surpassing even the supervised model (3.11/2.60): VL-T5$_{\mathrm{HatReD}}$ is fine-tuned on the annotation references alone, without access to the sociocultural knowledge that underlies each metaphor, so its outputs reflect limited metaphor-understanding capability, whereas knowledge-grounded rationales are judged more informative.

4.4 Quality of Quadrant Counterparts (RQ3)

From the 34,000 quadrant counterparts of the FHM training set, we draw a fixed-seed random sample of 200 (50 per quadrant, stratified by the source meme's polarity) and assess it with (i) 5-point Likert ratings, from MLLM and human raters, of the relevance between each rendered counterpart

Table 2: BERTScore of generated metaphor rationales against HatReD human annotations; best training-free results in bold

Method	Training-Free	BERT-P	BERT-R	BERT-F
VL-T5$_{\text{HatReD}}$ (sup.)	×	0.432	0.409	0.420
Baseline	✓	0.209	0.280	0.239
Baseline+L	✓	0.221	0.300	0.255
Baseline+L+S	✓	0.252	0.351	0.293
Baseline+L+S+G (Ours)	✓	**0.369**	**0.399**	**0.384**

Table 3: Likert ratings (1–5) of the content relevance between generated rationales and human annotations

Method	Training-Free	MLLM Panel	Human Panel
VL-T5$_{\text{HatReD}}$ (sup.)	×	3.112	2.602
Baseline	✓	3.614	2.931
Baseline+L	✓	3.817	3.283
Baseline+L+S	✓	4.029	3.850
Baseline+L+S+G (Ours)	✓	**4.882**	**4.040**

Table 4: Likert ratings (1–5) of the content relevance between the rendered quadrant counterparts and the required metaphor polarity

Quadrant	MLLM Panel	Human Panel
Q_1 (similar, consistent)	3.946	4.344
Q_2 (dissimilar, consistent)	3.943	4.467
Q_3 (dissimilar, opposite)	3.929	4.426
Q_4 (similar, opposite)	3.976	4.540

Table 5: CLIP cosine similarity ($\times 100$), computed with CLIP ViT-B/32, between the synthesized images and the image prompts

Quadrant	Mean	Median	Max
Q_1 (similar, consistent)	35.154	35.192	48.559
Q_2 (dissimilar, consistent)	35.641	35.608	48.369
Q_3 (dissimilar, opposite)	34.801	34.847	50.314
Q_4 (similar, opposite)	35.022	34.903	47.391

(the synthesized image with its overlaid text) and the required metaphor polarity of its quadrant (Table 4; protocol in Appendix B); (ii) CLIP cosine similarity Hessel et al. (2021) is additionally computed between every synthesized image and its image prompt over the full set (Table 5). Both panels rate all four quadrants highly and almost uniformly (MLLM: 3.93–3.98; human: 4.34–4.54), showing that the counterparts faithfully express the required polarity and that flipping it does not compromise label faithfulness. Mean CLIP cosine similarities Hessel et al. (2021) lie in a narrow band of 34.8–35.6 ($\times 100$) with medians almost identical to the means, matching the 30–35 range typical of state-of-the-art text-to-image models on short prompts; since the image prompts are concise natural-language descriptions, unlike the long rationale, this band indicates that text-free synthesis stays aligned with the prompts without quadrant-specific bias. Beyond image–prompt alignment, the surface-similarity dimension itself is verified in Appendix E: the image similarity between each source image and its similar-surface counterparts (Q_1, Q_4) is substantially higher than for its dissimilar-surface counterparts (Q_2, Q_3).

4.5 Ablation Study (RQ4)

Table 6 ablates the counterparts and QCL. Used purely as extra cross-entropy data, the counterparts lift FHM from 71.3/66.8 to 76.7/74.7 ACC/F1, with the confusion quadrants (Q2, Q4) contributing most of the F1 gain; replacing their cross-entropy supervision with QCL (Eqs. equation 11–equation 12) on the same data adds a further $+4.9$ ACC/$+6.5$ F1 on FHM ($+2.4$/$+2.5$ ACC on Harm-C/Harm-P), so on FHM, roughly half of the total gain comes from the counterparts as data and half from arranging them by metaphor polarity. The bottom block isolates the three QCL terms: removing the repulsion term hurts most on confounding-dense FHM (78.0/75.6), falling back toward pure CE augmentation; removing the regression term causes the largest drop on Harm-P (F1 88.5, the only cell below CE augmentation), where harmfulness hinges on entity-specific context and binding the polarity-consistent region outweighs expelling confusers; the ordering term contributes the least, as its constraints largely overlap the regression target gap (1.0–0.9).

4.6 Stability and Significance (RQ5)

Across ten random seeds (Table 7; RGCL re-run by us under the same splits and seeds, with results consistent with its published numbers), MQSCL is stable (std ≤ 0.15 ACC versus up to 0.65 for RGCL), suggesting the quadrant constraints also regularize optimization, and paired t-tests confirm significant improvements on all three datasets ($p < 10^{-8}$).

4.7 Case Study (RQ6)

Figure 3 examines two source memes. The rationale grounds the metaphor in external sociocultural knowledge (e.g., reading the Muslim kid's backpack as "the bomb") rather than restating visible content, and Q_4 shares the anchor's surface yet flips its stance—the exact confounder motivating MQSCL. Before QCL, similarity tracks surface and ranks the confounded pair on top; after QCL

Table 6: Ablation study on FHM, Harm-C, and Harm-P; the bottom block ablates the three calibration terms of QCL

Configuration	FHM ACC	FHM F1	Harm-C ACC	Harm-C F1	Harm-P ACC	Harm-P F1
Baseline (CE only)	71.3	66.8	85.9	81.3	85.4	85.2
+ Q2, Q4 (CE augmentation)	74.2	73.0	86.4	81.4	86.7	86.4
+ Q1–Q4 (CE augmentation)	76.7	74.7	87.2	82.8	88.9	88.8
+ Q1–Q4 + QCL (Ours)	**81.6**	**81.2**	**89.6**	**87.1**	**91.4**	**90.7**
+ Q1–Q4 + QCL w/o $\mathcal{L}_{\text{reg}}$	78.9	78.7	87.6	85.3	89.4	88.5
+ Q1–Q4 + QCL w/o $\mathcal{L}_{\text{rep}}$	78.0	75.6	87.7	83.4	89.4	89.0
+ Q1–Q4 + QCL w/o $\mathcal{L}_{\text{ord}}$	80.2	80.0	89.3	86.7	91.1	89.3

Table 7: Randomness analysis: mean±std of ACC over ten seeds, and paired t-tests against the RGCL re-runs.

Dataset	RGCL (re-run)	MQSCL (Ours)	t	p
FHM	78.81±0.27	**81.63±0.15**	39.60	2.1×10^{-11}
Harm-C	85.55±0.43	**89.57±0.08**	34.42	7.1×10^{-11}
Harm-P	86.65±0.65	**91.36±0.13**	20.79	6.3×10^{-9}

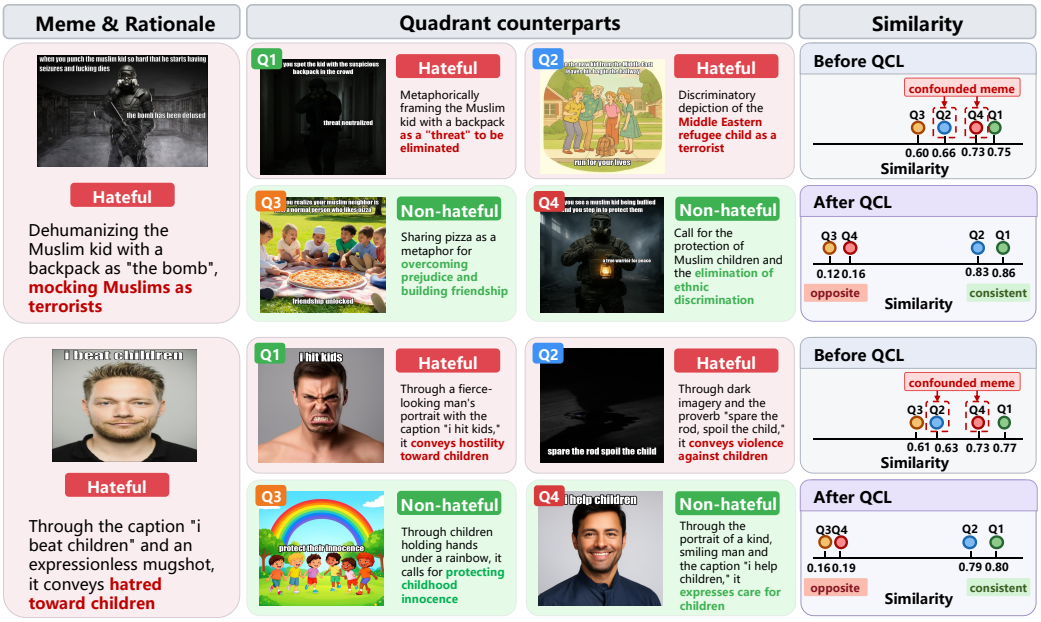

Figure 3: Case study: for two source memes, the metaphor rationale from the multi-agent analysis, the four quadrant counterparts, and anchor–counterpart similarities before vs. after QCL (readouts normalized to $[0, 1]$ for display).

it tracks stance, following $s_{Q_1} > s_{Q_2} > s_{Q_4} > s_{Q_3}$ (0.86/0.81 vs. 0.26/0.17 and 0.80/0.74 vs. 0.29/0.16 for the two memes), which expels the near-identical confounder from the anchor's neighborhood. The similarity readouts are normalized to $[0, 1]$ for display; in raw cosine space, the opposite-polarity counterparts of both memes fall below zero, satisfying the repulsion bounds of Eq. equation 11 ($s_{Q_4} \leq 0.1$, $s_{Q_3} \leq 0$).

5 Conclusion

Hateful meme detection is confounded by meme pairs that share almost the same surface yet convey opposite metaphors. MQSCL reduces this problem to a four-quadrant structure along surface similarity and metaphor-polarity consistency: multi-agent rationale analysis grounds metaphor understanding, four synthetic counterparts per source meme realize the quadrants, and the Quadrant Calibration Loss (QCL) reorganizes the feature space so that the decision boundary follows metaphor polarity rather than surface similarity. The counterparts are training-time scaffolding with no inference cost. On FHM, Harm-C, and Harm-P, MQSCL outperforms the strongest existing detectors (on FHM by 3.9 F1 over RGCL) with stable, statistically significant gains. As a general recipe for injecting the contrastive structure that real data rarely provides, the quadrant formulation should benefit other confusable multimodal categories.

REFERENCES

Bikramaditya Barman, Md Mukutor Rahman, and Rashel Sarkar. Gpt-5.5: evolution of agentic artificial intelligence (ai) in libraries. *Library Hi Tech News*, pp. 1–4, 2026.

Rui Cao, Roy Ka-Wei Lee, Wen-Haw Chong, and Jing Jiang. Prompting for multimodal hateful meme classification. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 321–332, 2022.

Rui Cao, Ming Shan Hee, Adriel Kuek, Wen-Haw Chong, Roy Ka-Wei Lee, and Jing Jiang. Procap: Leveraging a frozen vision-language model for hateful meme detection. In *Proceedings of the 31st ACM International Conference on Multimedia*, pp. 5244–5252, 2023.

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

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Ming Shan Hee and Roy Ka-Wei Lee. Demystifying hateful content: Leveraging large multimodal models for hateful meme detection with explainable decisions. *arXiv preprint arXiv:2502.11073*, 2025.

Ming Shan Hee, Roy Ka-Wei Lee, and Wen-Haw Chong. On explaining multimodal hateful meme detection models. In *Proceedings of the ACM Web Conference 2022*, pp. 3651–3655, 2022.

Ming Shan Hee, Wen-Haw Chong, and Roy Ka-Wei Lee. Decoding the underlying meaning of multimodal hateful memes. *arXiv preprint arXiv:2305.17678*, 2023.

Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. CLIPScore: A reference-free evaluation metric for image captioning. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih (eds.), *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 7514–7528, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.595. URL https://aclanthology.org/2021.emnlp-main.595/.

Douwe Kiela, Suvrat Bhooshan, Hamed Firooz, Ethan Perez, and Davide Testuggine. Supervised multimodal bitransformers for classifying images and text. *arXiv preprint arXiv:1909.02950*, 2019.

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

Gokul Karthik Kumar and Karthik Nandakumar. Hate-clipper: Multimodal hateful meme classification based on cross-modal interaction of clip features. In *Proceedings of the second workshop on NLP for positive impact (NLP4PI)*, pp. 171–183, 2022.

Black Forest Labs. FLUX.2: Frontier Visual Intelligence. https://bfl.ai/blog/flux-2, 2025.

Jian Lang, Rongpei Hong, Ting Zhong, Leiting Chen, Qiang Gao, and Fan Zhou. From shallow humor to metaphor: Towards label-free harmful meme detection via lmm agent self-improvement. In *Proceedings of the 32nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 1*, pp. 588–599, 2026.

Hongzhan Lin, Ziyang Luo, Jing Ma, and Long Chen. Beneath the surface: Unveiling harmful memes with multimodal reasoning distilled from large language models. *arXiv preprint arXiv:2312.05434*, 2023.

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

Phillip Lippe, Nithin Holla, Shantanu Chandra, Santhosh Rajamanickam, Georgios Antoniou, Ekaterina Shutova, and Helen Yannakoudakis. A multimodal framework for the detection of hateful memes. *arXiv preprint arXiv:2012.12871*, 2020.

Junyu Lu, Bo Xu, Xiaokun Zhang, Haohao Zhu, Kaichun Wang, Liang Yang, and Hongfei Lin. Is having rationales enough? rethinking knowledge enhancement for multimodal hateful meme detection. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 559–569, 2025.

Jingbiao Mei, Jinghong Chen, Weizhe Lin, Bill Byrne, and Marcus Tomalin. 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)*, pp. 5333–5347, 2024.

Niklas Muennighoff. Vilio: State-of-the-art visio-linguistic models applied to hateful memes. *arXiv preprint arXiv:2012.07788*, 2020.

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

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

Shivam Sharma, Md Shad Akhtar, Preslav Nakov, and Tanmoy Chakraborty. Disarm: Detecting the victims targeted by harmful memes. *arXiv preprint arXiv:2205.05738*, 2022.

Xuanyu Su, Yansong Li, Diana Inkpen, and Nathalie Japkowicz. A context-aware contrastive learning framework for hateful meme detection and segmentation. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pp. 5201–5215, 2025.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, et al. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*, 2025.

Qwen Team. Qwen3.5-omni technical report. *arXiv preprint arXiv:2604.15804*, 2026.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Chuanpeng Yang, Fuqing Zhu, Guihua Liu, Jizhong Han, and Songlin Hu. Multimodal hate speech detection via cross-domain knowledge transfer. In *Proceedings of the 30th ACM International Conference on Multimedia*, pp. 4505–4514, 2022.

Pengfei Yang, Haoyu Yang, Yili Li, Yong Wang, Jin Wu, Kunpeng Zhang, and Fan Zhou. Capsra: Lmm-guided caption semantic retrieval aggregation for hateful meme detection. *Pattern Recognition*, pp. 113781, 2026.

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

Xin Zhang, Yanzhao Zhang, Wen Xie, Mingxin Li, Ziqi Dai, Dingkun Long, Pengjun Xie, Meishan Zhang, Wenjie Li, and Min Zhang. Gme: Improving universal multimodal retrieval by multimodal llms. *arXiv preprint arXiv:2412.16855*, 2024.

APPENDIX

A DATASET STATISTICS

Table 8 summarizes the scale and label distribution of the three benchmarks. For every training meme in each dataset, the pipeline in Section 3.2 synthesizes four quadrant counterparts, one per quadrant. FHM additionally uses a separate 500-meme validation set (250 hateful / 250 non-hateful, shown in the table); Harm-C and Harm-P follow the train/test splits shown in the table and use no separate validation set.

Table 8: Statistical distributions of the datasets used for evaluation, and the number of generated quadrant counterparts

	Train		Test			
Dataset	**Hate**	**Non-hate**	**Hate**	**Non-hate**	**Total**	**Counterparts**
FHM Kiela et al. (2020)	3,050	5,450	500	500	9,500	34,000
Harm-C Pramanick et al. (2021a)	1,064	1,949	124	230	3,367	12,052
Harm-P Pramanick et al. (2021b)	1,486	1,534	173	182	3,375	12,080

B EVALUATION PROTOCOL

All subjective evaluations use the 5-point Likert scale Hee et al. (2023), where 1 means "not relevant at all" and 5 means "highly relevant"; each reported score is the mean over all ratings of the corresponding panel.

RQ2: rationale quality. Five third-party MLLM judges (Grok 4.6, Qwen-3.5-Plus, GPT-5.5, Claude Opus 4.8, and Gemini-3.5-Flash) and five human experts independently rate, on a random sample of 200 rationale–reference pairs drawn from the 3,228 memes covered by HatReD (fixed random seed), the content relevance between each generated metaphor rationale and its human reference. Each panel member independently rates all 200 samples (1,000 ratings per panel). The five human experts hold degrees in electronic information engineering (PhD), medicine (MSc), fine arts (MSc), geology (PhD), and communication studies (MSc).

RQ3: counterpart quality. The same MLLM and human panels as RQ2 rate the relevance between each rendered quadrant counterpart (the synthesized image with its overlaid text) and the required metaphor polarity of its quadrant on a random sample of 200 counterparts drawn from all $4 \times 8{,}500$ FHM training counterparts (50 per quadrant, stratified by the source meme's polarity; fixed random seed). Each panel member independently rates all 200 samples (1,000 ratings per panel). Raters see only the rendered counterpart and the required polarity of its quadrant, not the source meme or the generating model.

CLIP cosine similarity. The image-prompt alignment of every synthesized background image is scored with CLIP ViT-B/32 (`openai/clip-vit-base-patch32`); the score is the image–text cosine similarity clamped at zero and multiplied by 100. BERTScore in RQ2 is computed with `lang=en` and baseline rescaling, taking the best-scoring reference when multiple human annotations exist for a meme. The full text of the Likert evaluation prompt used by the MLLM panel is given in Appendix D.

C IMPLEMENTATION DETAILS

Table 9 summarizes the architecture and optimization hyperparameters of MQSCL.

Generation settings. The three rationale-analysis agents and the quadrant prompt agent all run on Gemini-3-Flash: the rationale-analysis agents use temperature 0.0 with JSON output and up to 3 retries with backoff, and the knowledge-retrieval call additionally enables the Google Search grounding tool; the quadrant prompt agent uses temperature 0.2 with JSON-mode output. Background images are synthesized with FLUX.2 Klein (base-4B UNet, Qwen3-4B text encoder, FLUX.2 VAE)

Table 9: Architecture and optimization hyperparameters of MQSCL

Module	Item	Setting
Feature extraction	Frozen multimodal encoder Φ	GME-2B
	Image/text embedding dim	1536
Fusion encoder $\mathcal{T}$	Input tokens (with modality markers, input LayerNorm)	[CLS] + [IMG] + [TXT]
	Layers / hidden size	4-layer pre-norm Transformer / 1536
	Attention heads	16
	FFN width / activation	1536 / GELU
	Dropout	0.3
	Output representation	ℓ_2-normalized [CLS] vector f
Classification head	MLP	2 layers (1536→512→2)
Optimization	Optimizer / peak learning rate	AdamW / 3×10^{-4}
	Learning-rate schedule	OneCycle (warmup fraction 0.1)
	Weight decay / gradient clipping	0.01 / 1.0
	Epochs	40
	Batch size	32
QCL	Loss weight λ	0.5

through ComfyUI at 768×768 resolution with 20 sampling steps, CFG 5.0, the Euler sampler under the simple scheduler, and denoise strength 1.0; seeds follow a deterministic schedule (base seed + sample index $\times$ 10 + quadrant index). Every positive image prompt is appended with the suffix "no text, no captions, no typography, no letters, clean background image only", and the full negative prompt is: "text, words, caption, meme text, typography, letters, watermark, logo, low quality, blurry, distorted face, bad anatomy, extra fingers, malformed hands, duplicate person, deformed body, wrong facial features, messy composition, misspelled text, unreadable text, cartoon, anime, painting, oversaturated, unrealistic lighting". The overlaid text is rendered from the normalized-coordinate text blocks produced by the quadrant prompt agent.

Hardware. All model training and evaluation run on a computing platform equipped with an NVIDIA 4090 GPU, a 128-core AMD EPYC 9754 CPU, 256 GB of memory, and 1 TB of disk storage.

Localization and filtering. Target localization (Eq. equation 3) and knowledge filtering (Eq. equation 4) are realized as in-context reasoning steps of the rationale generation agent under prompt p_3: the agent identifies relation paths that connect meme entities to group nodes and retains the knowledge items supported by these paths, without explicit numerical scoring or thresholding.

D Prompt Templates

This section lists the full text of the four pipeline prompts referenced in the main text (p_1: content interpretation and knowledge retrieval, run with the Google Search grounding tool; p_2: metaphor reasoning graph construction; p_3: target localization and rationale generation; p_4: quadrant generation specification), followed by the Likert evaluation prompt used by the MLLM panel in RQ2 and RQ3. Placeholders in angle brackets are filled with the corresponding meme fields at run time. Line breaks are reflowed for layout; the wording is otherwise as used at run time.

p_1: Content interpretation and knowledge retrieval

```
You are an expert in meme interpretation and social-cultural background
summarization. Please retrieve and summarize relevant social-cultural
knowledge based on the provided meme image and meme text, following the
rules below.

Meme information:
1. Meme text: <meme text>
2. Metaphor category (hateful / non-hateful): <metaphor label>
```

```
Rules:
1. Identify the core visual subject in the image and its symbolic attributes;
2. Extract named entities from the text, especially groups, events,
   locations, organizations, and historical or cultural symbols;
3. Retrieve relevant social-cultural knowledge based on the named entities
   in the text and the visual subject in the image, including social,
   cultural, historical, internet-related, or public-event knowledge;
4. The social-cultural knowledge must have clear and traceable reference
   sources, including authoritative webpages, news reports, encyclopedic
   entries, academic literature, or dataset documentation. Do not fabricate
   reference URLs, literature titles, or sources.

Please output the retrieved information in the following standard JSON
format:
{
  "related_elements": ["List of related persons / groups / events / symbols"],
  "key_summary": [
    {
      "knowledge": "Key social-cultural knowledge 1",
      "source": "Reference source"
    },
    {
      "knowledge": "Key social-cultural knowledge 2",
      "source": "Reference source"
    }
  ]
}
```

p_2: METAPHOR REASONING GRAPH CONSTRUCTION

```
You are an expert in graph-structured modeling. Please construct a
relational graph that can support large language model reasoning based on
the provided social background knowledge, following the rules below.

Social background knowledge: <retrieved knowledge>

Rules:
1. Node types may include the following categories:
- person: individuals related to the social background;
- group: social groups or online communities being referred to;
- event: relevant historical events, public events, social controversies,
  or online events;
- symbol: flags, logos, colors, abbreviations, slogans, patterns, or
  cultural symbols with symbolic meaning.
2. Edge relation types may include the following categories:
- belongs_to: indicates that a person belongs to a certain group;
- refers_to: indicates that a symbol points to or represents a group,
  event, or person;
- involved_in: indicates that a person or group participates in an event,
  causes an event to occur, or is affected by an event;
- derives_from: indicates that a node evolves from, is adapted from, or
  references another node.
3. The construction of nodes and edges must be supported by clear evidence
   from the provided social-cultural knowledge. Do not add nodes or edges
   that lack evidential support.
4. Each edge should provide a brief explanation of the evidence supporting
   the relationship.
5. Retain only nodes directly related to social-cultural reasoning, and
   ignore weakly related nodes or nodes that do not affect the core judgment.

Please output the result in standard JSON format according to the following
structure:
{
```

```
  "nodes": [
    {
      "id": "N1",
      "type": "person",
      "label": "Node name",
      "description": "Node description"
    }
  ],
  "edges": [
    {
      "source": "N1",
      "relation": "belongs_to",
      "target": "N2",
      "evidence": "Supporting evidence"
    }
  ]
}
```

p_3: TARGET LOCALIZATION AND RATIONALE GENERATION

```
You are an experienced expert in meme metaphor analysis. Please strictly
follow the rules below and generate the target group and metaphorical
reason of the meme based on the provided meme information and relational
graph.

Meme information:
1. Meme text: <meme text>
2. Metaphor category (hateful / non-hateful): <metaphor label>

Relational graph: <reasoning graph>

The graph contains the following node types:
- person: a person related to social background knowledge;
- group: a social group or online community being targeted or referred to;
- event: a relevant historical event, public event, social controversy, or
  online event;
- symbol: a flag, logo, color, abbreviation, slogan, pattern, or cultural
  symbol with symbolic meaning.

The graph contains the following edge types:
- belongs_to: indicates that a person belongs to a certain group;
- refers_to: indicates that a symbol points to or represents a group,
  event, or person;
- involved_in: indicates that a person or group participates in, causes,
  or is affected by an event;
- derives_from: indicates that a node evolves from, is adapted from, or
  references another node.

Rules:
1. Identify the target group by tracing relation paths from person, symbol,
   and event nodes to group nodes. Prioritize the group most closely
   associated with specific persons, symbols, and social events. Person
   nodes are usually used as supporting evidence for identifying the
   target group.
2. Around the target group, extract relevant socio-cultural knowledge
   based on its relation paths and supporting evidence, and convert it into
   concise semantic expressions.
3. Use the meme image and meme text as the primary basis, treat the given
   metaphor category as a constraint, and use the extracted socio-cultural
   knowledge as reference information to generate the corresponding
   metaphorical reason.
4. The reasonings should follow a concise dataset-style English sentence
   pattern, preferably using the structure: "[verb] [target group] by
   suggesting/portraying/implying/likening/framing them as [negative
```

```
   stereotype or metaphorical meaning]."

Please output the result in the following format:

Core reasoning path:
[{
    "path": "The core reasoning path obtained from the relational graph
             according to the defined rules",
    "meaning": "The meaning represented by the core reasoning path"
}]

target:
[A list of target groups. If there is no clearly supported hateful target,
output an empty list [].]

reasonings:
[A concise dataset-style sentence of no more than 31 words explaining why
the meme is hateful or non-hateful. Do not mention graph nodes or internal
reasoning processes.]
```

p_4: QUADRANT GENERATION SPECIFICATION

```
You are an expert in meme generation and metaphor analysis. Your task is to
construct meme generation plans for four quadrants based on the given meme
image, meme text, and metaphorical reasoning.

Input information:
1. Original meme text: <meme text>
2. Original metaphorical reasoning: <metaphor rationale>
3. Original meme image: provided as an attached image.

Quadrant definitions:
- Q1: Image-text features are similar, and the metaphorical meaning is
  consistent.
- Q2: Image-text features are dissimilar, and the metaphorical meaning is
  consistent.
- Q3: Image-text features are dissimilar, and the metaphorical meaning is
  opposite.
- Q4: Image-text features are similar, and the metaphorical meaning is
  opposite.

Please follow these rules:
1. First infer the original meme's visual style, subject type, scene,
   composition, and text layout based on the original meme image, meme
   text, and metaphorical reasoning.
2. "Similar image-text features" means preserving the original meme's
   layout structure, subject type, visual style, text placement, and
   sentence style as much as possible.
3. "Dissimilar image-text features" means changing the original meme's
   layout structure, subject type, visual style, text placement, or
   sentence style.
4. "Consistent metaphorical meaning" means that the new sample's core
   metaphorical meaning is basically consistent with the original
   metaphorical reasoning.
5. "Opposite metaphorical meaning" means that the new sample's core
   metaphorical meaning is opposite to the original metaphorical reasoning.
6. The image_prompt should describe only the background image content to be
   generated. Do not ask the image generation model to directly render meme
   text in the image_prompt.
7. Meme text must be placed in text_blocks so that it can be overlaid
   later using another post-processing program.
8. Text positions must use normalized coordinates in the range [0, 1].
   x_center and y_center indicate the center point of the text box. width
   and height indicate the proportion of the text box relative to the
```

```
   entire image.
9. Text should be concise and follow common meme style.
10. Output must be valid JSON only. Do not output explanations, Markdown,
    or code fences.

Please strictly follow this JSON format (Q2, Q3, and Q4 follow the same
schema with their own quadrant_name, feature_similarity, and
metaphor_relation fields):
{
  "original_layout_analysis": {
    "visual_style": "...",
    "main_subject": "...",
    "scene": "...",
    "composition": "...",
    "text_layout_summary": "..."
  },
  "quadrants": [
    {
      "quadrant_name": "Q1_similar_features_consistent_metaphor",
      "feature_similarity": "similar",
      "metaphor_relation": "consistent",
      "image_prompt": "...",
      "text_blocks": [
        {
          "text": "...",
          "position": {
            "x_center": 0.50,
            "y_center": 0.08,
            "width": 0.85,
            "height": 0.12
          }
        }
      ]
    }
  ]
}
```

LIKERT EVALUATION PROMPT (RQ2 AND RQ3)

The MLLM panel uses the following template; the evaluated content is the generated metaphor rationale together with its human reference in RQ2, and the rendered quadrant counterpart together with the required metaphor polarity of its quadrant in RQ3.

```
You are a meme metaphor analysis expert and a professional evaluator. Your
task is to evaluate the quality of the model-generated comprehensive
analysis of the input meme's image and text, based on the provided meme
information and scoring rules.

Meme information:
1. Meme text: <meme text>
2. Metaphor category (hateful / non-hateful): <metaphor label>

Model output:
<evaluated content>

Scoring rules:
1. Use a 5-point Likert scale to rate the relevance of the model-generated
   comprehensive analysis of the input meme's image and text.
- 1 point: Completely deviates from the meme meaning or provides a clearly
  incorrect explanation;
- 2 points: Partially relevant, but the explanation of the implicit meaning
  or hateful mechanism is clearly inaccurate;
- 3 points: Generally relevant, but the explanation is vague and contains
  some omissions or deviations;
```

```
- 4 points: Relatively accurate and explains the main metaphorical meaning
  and hateful/non-hateful judgment;
- 5 points: Accurately and completely explains the metaphorical meaning and
  hateful/non-hateful mechanism, with almost no obvious omissions.

2. The score must be an integer from 1 to 5.
3. The scoring reason should briefly explain how well the model output
   matches the meme meaning.
4. When scoring, consider the meme text, image content, and relevant
   social-cultural background to judge whether the model accurately
   explains the target, metaphorical meaning, and hateful/non-hateful
   mechanism. Give 1 point only when the output seriously deviates from
   the meme meaning or makes clearly incorrect attributions. Give 5 points
   only when the explanation is accurate, complete, highly consistent with
   the social context, and has almost no obvious omissions.

Output format:
{
  "score": <1-5>,
  "reason": "Give the reason for this score"
}
```

E VERIFICATION OF THE SURFACE-SIMILARITY MANIPULATION

The quadrant specification imposes a second, polarity-independent requirement: the counterparts of Q_1 and Q_4 must keep a surface similar to the source meme, while those of Q_2 and Q_3 must change it. CLIP cosine similarity (Table 5) verifies only image–prompt alignment and is blind to this dimension, so we additionally compute the cosine similarity between the frozen image embeddings of each source meme image v and its four counterpart images $\hat{v}_q$, using a frozen multimodal embedding model GME Zhang et al. (2024), over all 8,500 FHM training memes.

Table 10: Cosine similarity ($\times 100$) between the frozen image embeddings of the source meme and each quadrant counterpart, computed over all 8,500 FHM training memes

Quadrant	**Mean**	**Median**	**Max**
Q_1 (similar, consistent)	76.41	76.88	97.92
Q_2 (dissimilar, consistent)	41.72	41.69	71.13
Q_3 (dissimilar, opposite)	33.46	33.08	61.89
Q_4 (similar, opposite)	76.61	77.10	95.30

A clear separation between the similar-surface quadrants $\{Q_1, Q_4\}$ (mean similarity 76.41 and 76.61) and the dissimilar-surface quadrants $\{Q_2, Q_3\}$ (mean similarity 41.72 and 33.46) confirms that generation follows the surface-similarity dimension of the quadrant specification; together with the polarity ratings of Table 4 and the image–prompt alignment of Table 5, this verifies both control dimensions of the quadrant framework.