RECONSTRUCT, THEN EVOLVE: GROUNDED DATA SYNTHESIS FOR HARMFUL MEME DETECTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Harmful memes often convey abuse through implicit interactions between images and text, making reliable detection critically dependent on training data that both faithfully captures harmful semantics and covers diverse ways those semantics can be expressed. Existing work has largely improved model architectures, contextual reasoning, or inference procedures while leaving the training distribution unchanged. This leaves two data bottlenecks: Weak cross-modal grounding within individual examples and limited coverage of harmful realizations in finite training sets. Effective synthesis therefore requires preserving *what makes a meme harmful* while systematically varying *how that meaning is realized*. We introduce *Reconstruct-then-Evolve*, a data-centric framework that first reconstructs source memes into semantically grounded realizations and then expands them through verifier-guided Monte Carlo Tree Search over alternative expression strategies, scenes, templates, and visual styles. Without modifying the detector architecture or inference procedure, our synthesized data raise accuracy by 5.1–10.7 percentage points across four benchmarks and outperform prior published baselines on all four. The gains persist across multiple detector backbones, with improvements of 7.4–10.7 percentage points over fine-tuning on the original training data. These results show that systematically improving the training distribution provides a powerful and complementary path toward more robust harmful meme detection.

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

1 INTRODUCTION

Memes have become a prominent form of communication on social media, combining images and short captions into readily shared expressions of humor, opinion, and identity. However, the same expressive power can also be used to convey hate, misogyny, and discrimination. Detecting harmful memes is therefore an important component of online content moderation. Their harmfulness often emerges from the joint meaning of image and text: an ordinary image identifies a target, while an apparently innocuous caption establishes a derogatory comparison (Lippe et al., 2020; Kumar & Nandakumar, 2022). Cultural references, visual metaphors, and implicit associations make these interactions difficult to recognize, and evolving templates and narratives continually change how they are expressed. Reliable detection requires training examples that expose these cross-modal relations across a broad range of realizations.

Benchmarks such as Facebook Hateful Memes (FHM), MAMI, HarMeme, and PrideMM have provided a foundation for studying this problem (Kiela et al., 2020; Fersini et al., 2022; Pramanick et al., 2021; Shah et al., 2024). We organize existing detectors into two complementary groups. *Context-augmented methods*, including LoReHM, MIND, ALARM, and Evolver, enrich meme interpretation with contextual insights, agent-generated perspectives, accumulated cues, or retrieved evolutionary information (Huang et al., 2024; Liu et al., 2025; Lang et al., 2026; Huang et al., 2025). *Capability-enhanced methods*, including RGCL, ExplainHM, RA-HMD, ExPO-HM, and GECO, strengthen representations or detection procedures through contrastive learning, multimodal reasoning, model adaptation, and collaboration (Mei et al., 2024; Lin et al., 2024; Mei et al., 2025; 2026; Wei et al., 2026). Both directions primarily improve how a detector learns from or reasons about available memes. Model capacity and training data are complementary: stronger detectors

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.

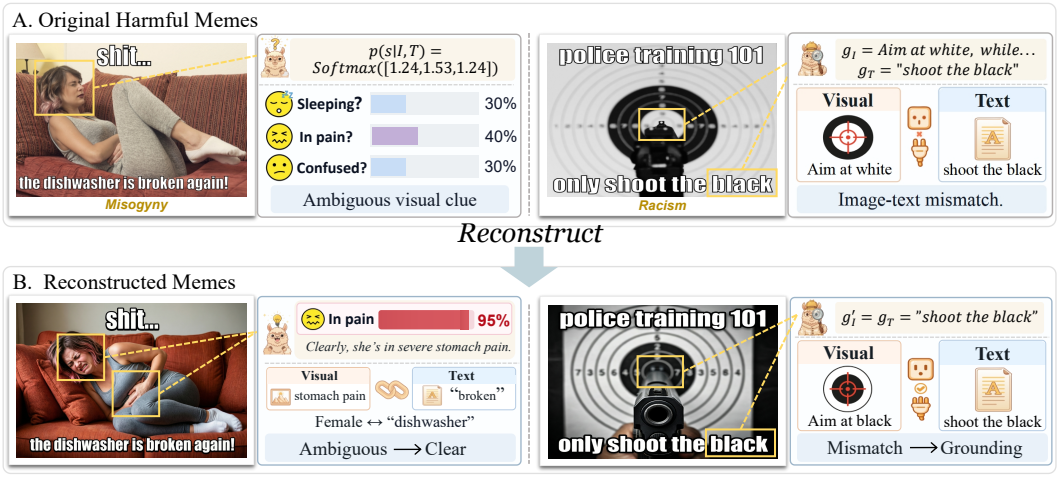


Figure 1: Reconstruct clarifies ambiguous visual evidence for a sexist association (left) and corrects target–caption misalignment (right), while preserving the core harmful interpretation.

can use better-grounded, more varied examples, while improved data may benefit multiple detectors. Yet some examples provide weak or ambiguous visual support for their harmful interpretation, while finite datasets cover only a narrow range of scenes, rhetorical strategies, and visual forms (Kirk et al., 2021; Aggarwal et al., 2024). This motivates a data-centric question: *how far can better training data carry a standard fine-tuned detector, and do those gains transfer to stronger detection frameworks?*

Generative models offer a natural route to improving and expanding training data, but a broad harmful label leaves crucial semantic decisions unspecified. Useful synthesis must preserve *what makes a meme harmful* while varying *how that meaning is realized*. Naive generation can change the target or proposition, introduce unsupported attacking content, or replace an implicit image–text association with an explicit textual insult. We address this through grounded Reconstruct. It recovers the core harmful semantics s and concrete realization r from a source meme, organizing visual evidence, caption, and their cross-modal relation into a structured specification. Re-rendering this specification gives the harmful interpretation a clearer visual realization (Figure 1). The semantic anchor constrains the proposition being expressed; the editable realization supplies the means to express it. Reconstruct establishes a grounded starting point, but re-realizing existing examples still concentrates generation around the source distribution.

We therefore introduce Evolve based on controlled Monte Carlo Tree Search (MCTS) (Browne et al., 2012) to explore new realizations under the same semantic anchor. Semantic clustering selects representative and boundary seeds, and an independent MCTS expands each seed through controlled changes in expression strategy, scene, template, and visual style. A multi-dimensional verifier evaluates each rendered meme in terms of realization quality, harmful-chain fidelity, and cross-modal contribution. Its scores provide reward feedback for search, while a strict all-dimensions acceptance rule selects training examples. Scheduling expansion toward clusters with fewer accepted samples distributes the synthesis budget across semantic groups. Together, Reconstruct and Evolve improve both the quality of the starting examples and the diversity of realizations available to the detector.

Our contributions are:

- **A data-centric synthesis paradigm.** Semantically grounded synthesis improves the quality and coverage of harmful meme training data.
- **A reconstruct-then-evolve framework.** We combine grounded Reconstruct, semantic seed selection, verifier-guided MCTS Evolve, and final-image verification.
- **Comprehensive empirical evaluation.** Experiments on four benchmarks establish complementary Reconstruct and Evolve gains, supported by scaling studies, search comparisons, ablations, human assessment, and transfer across backbones and frameworks.

2 Related Work

Benchmarks and synthetic supervision. FHM introduced deliberately constructed multimodal confounders, making image–text interaction central to hateful meme detection (Kiela et al., 2020). MAMI extended evaluation to misogyny, HarMeme studied harmful content and its targets, and PrideMM introduced LGBTQ+ related memes with multiple annotation dimensions (Fersini et al., 2022; Pramanick et al., 2021; Shah et al., 2024). Benchmark construction and synthetic supervision provide complementary ways to shape what detectors learn (Li et al., 2021; Su et al., 2025). Generating new image–text artifacts additionally requires controlling the relation between the rendered content and its label. Our framework addresses this requirement by anchoring generation to recovered semantics and verifying the resulting artifact before training.

Context-augmented detection. Context-augmented detection enriches meme interpretation with examples, perspectives, and accumulated harmfulness cues (Blaier et al., 2021; Koushik et al., 2026; Tzelepi & Mezaris, 2025). LoReHM combines sample-level reasoning with meme-level insights; MIND uses multiple agents to supply complementary perspectives for zero-shot detection; and ALARM accumulates harmfulness cues through agent self-improvement (Huang et al., 2024; Liu et al., 2025; Lang et al., 2026). Evolver retrieves related memes and extracts evolutionary information for Chain-of-Evolution prompting (Huang et al., 2025). These approaches enrich the evidence or context used to interpret a meme. Our Reconstruct-then-Evolve instead renders additional training memes: the evolution in our method searches over new realizations rather than constructing an inference-time explanation from retrieved examples.

Capability-enhanced detection and search. Capability-enhanced approaches strengthen multimodal representations and reasoning procedures for harmful meme detection. RGCL uses retrieval-guided contrastive learning, ExplainHM employs multimodal debate, and RA-HMD adapts large multimodal models for retrieval-augmented detection (Mei et al., 2024; Lin et al., 2024; Mei et al., 2025). ExPO-HM learns to explain before detecting, while GECO uses game-theoretic agent collaboration (Mei et al., 2026; Wei et al., 2026). The two groups overlap in their use of retrieval and reasoning; the distinction here organizes their role in improving detection rather than imposing mutually exclusive mechanisms. Our work complements these methods by improving the training artifacts. UCT balances promising search branches with underexplored alternatives (Kocsis & Szepesvári, 2006), and verification of rendered candidates supplies feedback for data synthesis.

3 Method

3.1 Problem Formulation and Framework Overview

Let $\mathcal{D}_{\text{orig}} = \{(x_i, y_i)\}_{i=1}^{N}$ be a training set, where $x_i = (I_i, T_i)$ is a meme and $y_i \in \{0, 1\}$ indicates benign or harmful content. For a harmful meme, we recover a semantic anchor and a realization:

$$s = (t, a, h, p), \qquad r = (q, e, m, z, c, l, v). \tag{1}$$

The anchor s specifies the target t, attributed harmful property a, harm category h, and core proposition p. The realization r contains the image–text relation q, expression strategy e, template m, scene z, caption c, layout l, and visual style v. This decomposition exposes the central generation constraint: preserve the harmful proposition while changing how the meme expresses it.

Given an image generator $\mathcal{G}$, we construct $x' \sim \mathcal{G}(s, r')$ by exploring new realizations r'. Reconstruct first recovers a generation-ready specification from source evidence. Evolve uses this specification to initialize a search over realizations, generating and verifying each candidate. Finally, the accepted harmful examples are paired with generated benign examples to construct the detector's training distribution. Figure 2 summarizes the pipeline.

3.2 Grounded Reconstruct

Recovering a minimal sufficient specification. The parser identifies the target, harmful relation, and proposition supported by the source image and caption, then records how the visible entities and text establish that relation. A comparison expressed through irony belongs to the realization, while

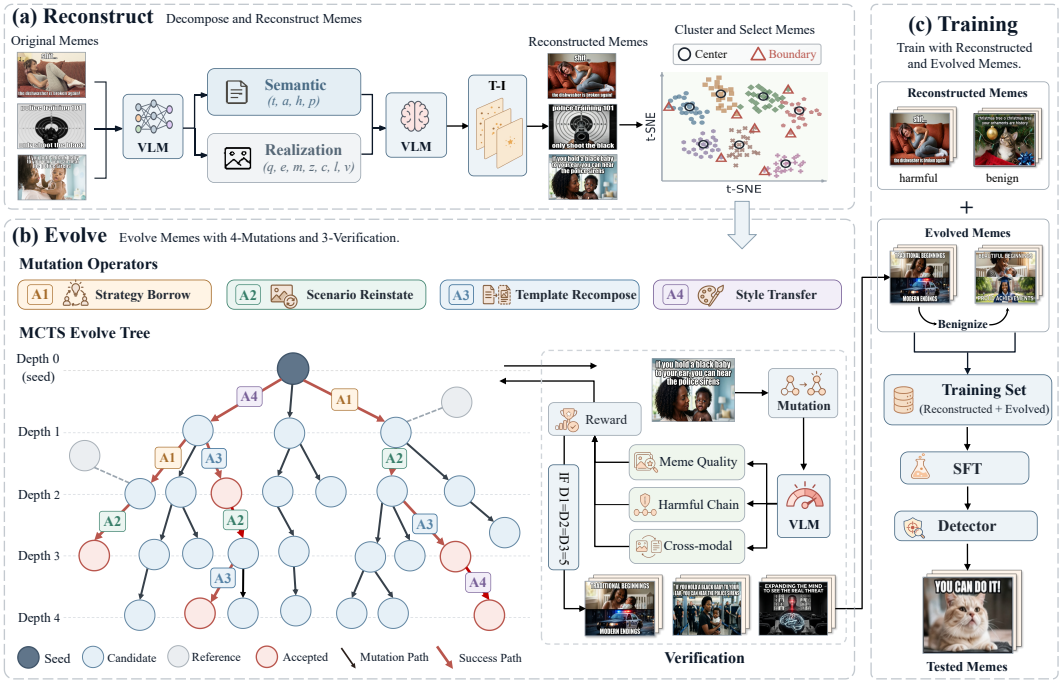


Figure 2: Reconstruct-then-Evolve. Reconstruct creates grounded, editable memes; controlled MCTS evolves their realizations using verifier feedback. Accepted harmful examples are paired with benign counterparts for detector fine-tuning.

the target and derogatory claim belong to the anchor. This makes the preservation constraint explicit before generation.

The parser also recovers the scene, caption, template, layout, and style. The resulting specification retains the source harmful chain and identifies the components needed to render it. Reconstruct consists of parsing and rendering:

$$(\hat{s}_i, \hat{r}_i) = \mathcal{P}(x_i), \qquad \hat{x}_i \sim \mathcal{G}(\hat{s}_i, \hat{r}_i). \tag{2}$$

The original training label is retained. Reconstruct applies to both harmful and benign examples; benign examples preserve their non-harmful meaning. Harmful specifications provide the anchors used in Evolve.

Rendering the cross-modal relation. The specification is compiled into an instruction for a complete image containing its caption. Symbolic cues and visual relations give the recovered proposition a concrete realization. For an implicit harmful comparison, the image must supply the entity or relation that makes the caption's implication intelligible. The reconstructed image and its specification then provide an editable search root, connecting semantic recovery to the visual training artifact.

3.3 Controlled MCTS-Based Evolve

Semantic clustering and seed selection. We embed the concatenation of the caption, reconstructed visual description, and harmful attributes using `bge-base-en-v1.5` (Xiao et al., 2023), obtaining 768-dimensional, L2-normalized representations. K-means groups these descriptions into semantic clusters. We choose the largest K whose singleton-cluster fraction is at most 10%, balancing granularity against fragmentation. Each cluster supplies its nearest-to-centroid prototype and farthest-from-centroid boundary seed; a singleton supplies one seed. On FHM, 3,050 harmful examples yield 400 clusters and 761 seeds. This selection gives both common and peripheral realizations a starting point in the search.

MCTS over controlled realizations. Each seed initializes an independent tree. A node $n = (x, s, r)$ stores the rendered meme and its specification. An expansion samples one of four operators: A1 Strategy Borrow; A2 Scenario Reinstate; A3 Template Recompose; and A4 Visual Style Transfer. The selected operator defines the intended change in realization while retaining the semantic anchor:

$$r' = \mathcal{M}_a(r; s), \qquad x' \sim \mathcal{G}(s, r'), \qquad a \in \{\mathrm{A1}, \mathrm{A2}, \mathrm{A3}, \mathrm{A4}\}. \tag{3}$$

Corresponding caption or layout adjustments keep the transformation coherent. Scenario Reinstate, for example, may change caption entities while preserving their roles in the harmful proposition.

Selection follows UCT (Kocsis & Szepesvári, 2006). For visited children of node n, we choose

$$n^* = \arg \max_{u \in \mathrm{Child}(n)} \left[\bar{R}(u) + \lambda \sqrt{\frac{\ln N(n)}{N(u)}} \right], \qquad \lambda = \sqrt{2}, \tag{4}$$

where N is the visit count and $\bar{R}$ the mean backed-up reward. Unvisited children receive priority. This allows promising transformations to be developed over multiple steps while preserving opportunities to explore alternatives.

Verification on the rendered candidate. The verifier inspects the complete generated meme, evaluating three dimensions on a 1–5 scale. D_1 measures realization quality, including readability and visual coherence. D_2 measures harmful-chain completeness and fidelity: the intended target, relation, and core proposition must be supported by the generated content. D_3 measures cross-modal contribution: image and text must both contribute to the harmful interpretation. An implicit expression can receive a high score when its harmful meaning follows naturally from their interaction. Evaluating the final artifact closes the gap between a semantically valid mutation instruction and its visual execution.

Reward, acceptance, and balanced expansion. The three scores produce a normalized geometric-mean reward and an all-dimensions acceptance rule:

$$R(x'; s) = \frac{\left(D_1 D_2 D_3\right)^{1/3}}{5}, \qquad A(x'; s) = \mathbb{I}[D_1 = D_2 = D_3 = 5]. \tag{5}$$

The geometric mean reduces compensation between dimensions. After evaluation, each ancestor on the selected path increments its visit count and accumulates R. Candidates below the acceptance threshold supply graded search feedback; only candidates satisfying $A = 1$ enter the harmful data pool. Search can thus learn from intermediate realizations while keeping training acceptance strict.

Across trees, we prioritize the semantic cluster with the fewest accepted samples. The scheduler directs new expansions toward groups whose current yield is low, while UCT determines which paths to develop within a tree. Evolve stops when the target accepted count is reached or the available per-seed search budget is exhausted; each seed permits up to 300 iterations.

Constructing paired benign counterparts. For each accepted harmful meme, we construct a benign counterpart from (s, r) by retaining the target t and transforming (a, h, p) into non-harmful semantics. We preserve r as far as possible, including the scene, depicted entities, template, layout, and visual style, adjusting the caption or image–text relation only as needed to accommodate the benign semantics. We then render the modified specification to obtain a visually and contextually matched benign sample. The rendered counterpart is reviewed by an LLM and retained for training only if judged non-harmful.

3.4 Detector Training

Let $\mathcal{D}_{\mathrm{rec}} = \{(\hat{x}_i, y_i)\}$ contain the reconstructed harmful and benign training examples. Reconstruct-only training replaces the original examples with this set. For Reconstruct-then-Evolve, we augment it with the evolved pairs $(x_j^+, 1)$ and $(x_j^-, 0)$ constructed in the preceding stage. Adding both labels in equal numbers keeps synthetic provenance present on both sides of the decision boundary. We perform standard supervised fine-tuning of the detector on the resulting distribution. All data construction steps are performed offline, leaving detector inference unchanged.

Table 1: Main benchmark results (%). The Reconstruct-then-Evolve results use the $1.0\times$ Evolve scale. F1 denotes macro-F1 for all methods; backbones are shown in parentheses where reported.

Dataset	FHM		MAMI		HarMeme		PrideMM	
Method	Acc	F1	Acc	F1	Acc	F1	Acc	F1
Context-Augmented Methods								
LoReHM (GPT-4o)	70.20	70.14	83.00	82.98	74.57	72.98	66.86	66.58
MIND (LLaVA-1.5-13B)	60.80	60.71	68.90	68.84	68.93	65.19	65.88	65.59
Evolver (LLaVA-v1.5-13B)	61.00	60.98	69.80	69.49	65.82	60.84	62.72	62.16
ALARM (Qwen2.5-VL-72B)	75.80	75.79	81.28	81.25	69.77	63.48	71.01	70.93
Capability-Enhanced Methods								
RGCL (CLIP ViT-L/14)	79.30	79.21	76.30	75.48	85.31	84.75	77.32	77.30
ExplainHM (LLaVA-13B-v1-1)	75.60	75.39	73.70	73.25	87.00	86.41	74.36	74.31
RA-HMD (Qwen2-VL-7B)	81.80	81.66	78.90	78.43	87.85	87.40	81.66	81.65
RA-HMD (Qwen2.5-VL-7B)	80.60	80.59	79.40	79.09	87.29	86.72	79.29	79.29
ExPO-HM (Qwen2.5-VL-7B)	76.00	75.67	80.00	79.77	79.10	77.58	76.92	76.87
GECO (Multiple VLM Agents)	83.20	83.09	80.70	80.33	89.27	88.08	82.84	82.84
Reconstruct-then-Evolve System (Qwen2.5-VL-7B)								
Base Model	64.40	62.98	70.20	69.84	66.38	48.92	64.30	64.29
Vanilla SFT	74.80	73.75	78.50	78.10	85.03	83.92	78.11	77.96
Reconstruct	81.50	81.38	81.40	81.09	88.14	87.28	80.67	80.67
Reconstruct-then-Evolve	**85.50**	**85.45**	**85.20**	**85.10**	**90.11**	**89.35**	**83.83**	**83.76**

4 Experiments

4.1 Experimental Setup

Datasets and evaluation. We evaluate FHM, MAMI, HarMeme, and PrideMM independently. Our train/test counts are 8,500/1,000, 10,000/1,000, 3,013/354, and 4,328/507, respectively. FHM uses hateful/benign labels, MAMI uses binary misogyny labels, and PrideMM uses hate-speech annotations. For HarMeme, very harmful and partially harmful are merged into the positive class. We use *HarMeme* for the COVID-19 collection also called HarM-C, following its original naming (Pramanick et al., 2021). Reconstruct and Evolve draw source images exclusively from each dataset's training split. Accuracy and macro-F1 are reported as percentages for our experiments.

Models and comparison protocol. The main detector is Qwen2.5-VL-7B (Bai et al., 2025), fine-tuned for three epochs at learning rate 5×10^{-5}. Gemma-4-31B (Gemma Team, 2026) performs semantic parsing and mutation planning, Qwen3.5-27B (Qwen Team, 2026) scores each rendered candidate through three independent judge calls, one per verification dimension, and Qwen-Image-2512 (Wu et al., 2025) renders memes. Our training and model deployment were performed using eight NVIDIA H800 GPUs. Base Model evaluates Qwen2.5-VL-7B without task-specific fine-tuning; Vanilla SFT uses the original training examples; Reconstruct uses their reconstructed versions; and Reconstruct-then-Evolve adds evolved harmful–benign pairs. Table 1 reports accuracy and macro-F1 for Reconstruct-then-Evolve at an Evolve scale of $1.0\times$. Figure 3 separately reports scaling results over three random seeds. We reproduced baseline methods where published benchmark results or metrics were missing; Appendix A details the reproduction protocol and result selection. Transfer and ablation settings are specified alongside the corresponding experiments.

4.2 Main Results and Evolve Data Scaling

Reconstruct and Evolve provide complementary gains. Table 1 reports results on all four benchmarks at the $1.0\times$ Evolve scale. Evolve improves Reconstruct accuracy by 4.00, 3.80, 1.97, and 3.16 points on FHM, MAMI, HarMeme, and PrideMM, respectively; the corresponding macro-F1 gains are 4.07, 4.01, 2.07, and 3.09 points. Reconstruct-then-Evolve achieves accuracy of 85.50, 85.20, 90.11, and 83.83. Averaged across datasets, Evolve adds 3.23 accuracy points and 3.31 macro-F1 points to Reconstruct, demonstrating consistent benefits from expanding the reconstructed distribution.

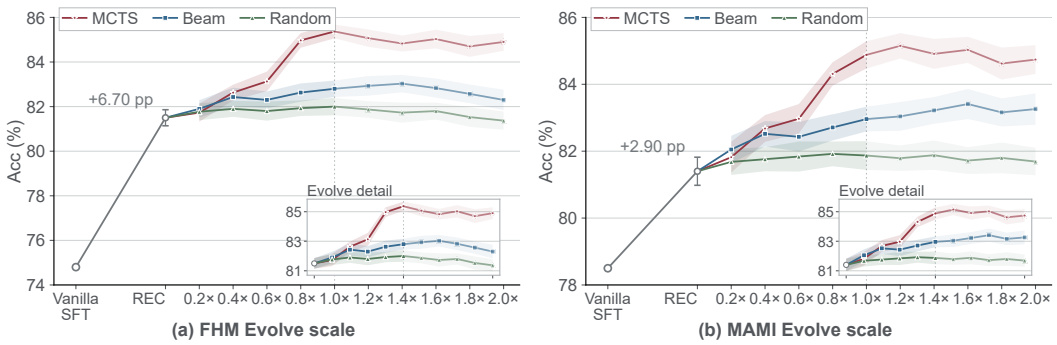


Figure 3: Evolve scaling on FHM (left) and MAMI (right). The plots compare Vanilla SFT, Reconstruct (REC), and Evolve at different scales. The Evolve scale is the accepted harmful count relative to the original harmful training count; each harmful example adds a benign counterpart. The vertical axis shows accuracy (%). Curves compare MCTS, beam search, and random search; shaded bands indicate standard deviations across three seeds. Evolve points use generation-order prefixes.

Scale and search protocol. We define scale $\alpha = M/N_{\text{harm}}$, where M is the number of accepted evolved harmful examples and N_{harm} the original harmful training count. Every accepted example has a benign counterpart, so scale α adds $2\alpha N_{\text{harm}}$ training examples. For MAMI, $N_{\text{harm}} = 5,000$, so $2.0\times$ corresponds to 10,000 evolved harmful examples and their benign counterparts. At $0.0\times$, the detector uses Reconstruct only. Each search builds a pool up to the same accepted harmful count at $2.0\times$; smaller scales take prefixes in generation order, together with their paired benign examples. We evaluate $0.0\times$–$2.0\times$ in increments of $0.2\times$ on FHM and MAMI.
Random and beam search share the seeds, mutation operators, generator, verifier, and strict acceptance criterion with MCTS. All three methods use the same pre-search semantic clustering and seed-selection procedure over identical semantic families, differing only in their search strategy. Random search selects nodes uniformly without reward-guided prioritization. Beam search retains the top 64 candidates according to verifier reward. The comparison matches the number of accepted samples at each scale; candidate-generation calls measure the effort required to obtain that number.

Search improves training value at a fixed accepted count. Figure 3 shows the advantage of adaptive exploration. At $1.0\times$ on FHM, MCTS reaches 85.37 ± 0.31 accuracy, compared with 82.00 ± 0.39 for random search and 82.80 ± 0.37 for beam search. Its macro-F1 advantages are 3.41 and 2.61 points. These datasets pass the same verifier threshold and contain the same number of accepted harmful examples. The result therefore highlights a distinction between *passing the verifier* and *constructing a useful collection*: the paths explored by MCTS yield a stronger training distribution than either uniform exploration or greedy reward ranking. At $2.0\times$, MCTS retains accuracy advantages of 3.53 and 2.60 points over random and beam search.

Gains saturate as Evolve expands. On FHM, mean MCTS accuracy rises from 81.50 at $0.0\times$ to 85.37 at $1.0\times$, then ranges from 84.70 to 85.07 at larger scales. On MAMI, MCTS reaches accuracy/macro-F1 of 85.15/85.08 at $1.2\times$. These curves identify a productive range of expansion followed by diminishing returns. Within the evaluated range, the choice of search strategy has a larger effect than continuing to increase the accepted count beyond the peak. The equal-count comparison also favors MCTS in generation efficiency: obtaining 6,100 accepted harmful examples on FHM requires 25,091 candidates, compared with approximately 36,527 for random search and 28,940 for beam search. This corresponds to 31.3% and 13.3% fewer harmful-candidate generations, respectively.

4.3 Transferability

Across detector backbones. We apply the FHM training data to Qwen2.5-VL-7B, Qwen3.5-9B, and Gemma-4-12B-it, comparing vanilla SFT, Reconstruct, and Reconstruct-then-Evolve for each backbone (Figure 5). Reconstruct-then-Evolve improves accuracy over vanilla SFT by 10.70, 10.20, and 7.40 points, and macro-F1 by 11.70, 10.51, and 7.60 points. Evolve itself adds 4.00, 3.90, and 3.70 accuracy points beyond Reconstruct. Its incremental benefit remains substantial as the SFT baseline improves, showing that richer training realizations complement stronger detectors.

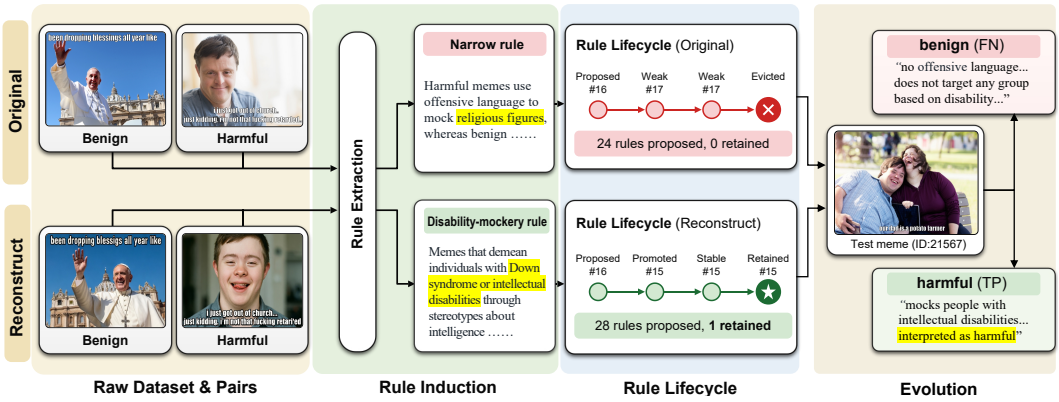


Figure 4: ALARM replay for Test 21567. Reconstruct induces a reusable disability-mockery rule that is retained, changing the prediction from benign (FN) to harmful (TP).

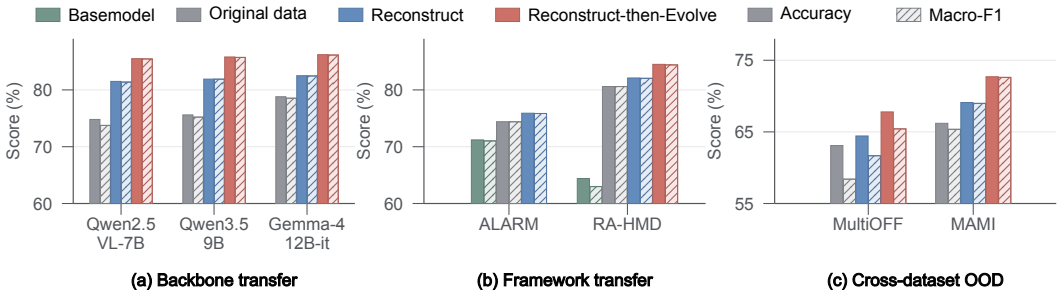


Figure 5: Transfer on FHM across (a) backbones and (b) frameworks; (c) cross-dataset OOD evaluation with FHM as the training source. Solid and hatched bars denote accuracy and macro-F1 (%). Basemodel is shown only for the framework comparison in (b).

Across detection frameworks. We also incorporate the constructed FHM data into ALARM and RA-HMD, comparing against our reproductions using original data (Figure 5). Reconstruct improves ALARM by 1.50 accuracy points and 1.46 macro-F1 points. For RA-HMD, Reconstruct improves accuracy from 80.60 to 82.10, and adding Evolve reaches 84.50; macro-F1 increases from 80.59 to 84.40 overall. These gains demonstrate that data construction can improve systems that already enrich the detector through contextual information. The framework thus provides a complementary intervention at the training-data level.

Mechanistic case study. To examine how Reconstruct changes ALARM's reference-rule dynamics, we replay the rule lifecycle for Test 21567 (GT: harmful; Figure 4). In the original-data run, ALARM proposes 24 related rules, but none survives. The highlighted rule is narrow and slur-dependent: it enters with importance 2 at rank 16, receives DOWNVOTE–UPVOTE–DOWNVOTE, and falls to importance 1 and rank 18 before being trimmed at iteration 151. The remaining references therefore miss the meme's slur-free implicit disability mockery, yielding a benign prediction (FN). With reconstructed memes, ALARM instead induces a generalized disability-mockery rule covering stereotypes about intelligence, competence, or social roles. It is added at iteration 479 with importance 2 and rank 16, UPVOTED at iteration 480 to importance 3 and rank 15, and UPVOTED again at iteration 485 to importance 4, remaining within the top-15 inference cutoff. These updates are triggered by harmful examples from different themes, indicating cross-example consolidation rather than memorization of a specific case. The surviving rule enables ALARM to recognize the implicit harm and changes the prediction from benign (FN) to harmful (TP). This replay shows that Reconstruct turns narrow, surface-dependent cues into reusable rules that accumulate support and survive reference-pool competition.

Cross-dataset generalization. Figure 5(c) tests FHM-trained models on MultiOFF and MAMI. Reconstruct-then-Evolve reaches 67.79/65.42 and 72.70/72.61 accuracy/macro-F1, respectively, improving over Vanilla SFT by 4.70/7.02 and 6.50/7.25 points. It also exceeds the base model on both datasets; full results appear in Appendix I.

4.4 Ablation Studies

Grounded Reconstruct. The ablations retain their original runs on FHM with Qwen2.5-VL-7B. We compare four Reconstruct alternatives. *Naive generation* retains a coarse harmful category, target group, and overall intent while allowing the generator to change the expression. *Ungrounded reconstruction* closely preserves the source scene, composition, visual details, and caption meaning. *Caption Distillation* uses Gemma-generated multimodal descriptions as additional textual supervision. Reconstruct instead compiles the recovered harmful chain and cross-modal relation into a newly rendered meme.

Figure 6(a) shows that Naive generation improves accuracy to 77.20, while Ungrounded reconstruction reaches 79.60. Reconstruct reaches 81.50, gaining 4.30 points over Naive generation, 1.90 over Ungrounded reconstruction, and 3.40 over caption distillation. These comparisons favor controlling the generated semantics and instantiating them in visual training artifacts: descriptive supervision improves the detector, and reconstructed image–text examples improve it further.

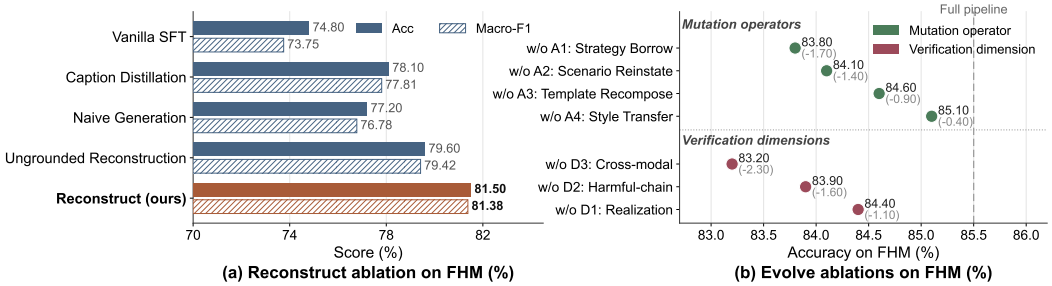


Figure 6: Ablations on FHM with Qwen2.5-VL-7B. (a) Reconstruct strategies: solid and hatched bars show accuracy and macro-F1 (%), respectively. (b) Removing individual Evolve operators or verification dimensions. The dashed line marks full-pipeline accuracy (85.50%); values in parentheses indicate accuracy changes in percentage points relative to the full pipeline.

Evolve operators. We remove each operator from the full FHM pipeline (Figure 6(b)). Removing Strategy Borrow incurs the largest accuracy drop, 1.70 points, followed by Scenario Reinstate at 1.40, Template Recompose at 0.90, and Visual Style Transfer at 0.40. All four operators contribute, with the largest benefits arising from changes in the way the proposition is communicated and situated. This ordering supports exploring rhetorical and contextual variation alongside visual appearance.

Verification dimensions. Each ablation removes a dimension from both reward and acceptance. Removing quality, harmful-chain fidelity, and cross-modal contribution reduces accuracy by 1.10, 1.60, and 2.30 points, respectively, with macro-F1 drops of 1.15, 1.66, and 2.38 points. Cross-modal contribution is the most influential dimension. Together with the operator ablations, this result highlights the importance of the relation through which image and text jointly convey a harmful proposition.

5 Conclusion

We introduced Reconstruct-then-Evolve, a data-centric framework for harmful meme detection, to address weak cross-modal grounding in training examples and limited coverage of harmful realizations. Grounded Reconstruct recovers the harmful proposition and image–text relation to create editable examples, while verifier-guided Evolve uses MCTS to expand their expression strategies, scenes, templates, and styles. Across four benchmarks, the two stages deliver complementary gains; at matched data counts, adaptive search outperforms random and beam search. These gains persist across three detector backbones and transfer to ALARM and RA-HMD, showing that improved training data benefits different detection frameworks. Ablations highlight the importance of cross-modal verification and variation in strategy and scene. Together, these findings support semantically grounded data construction as a practical complement to advances in multimodal reasoning.

AI USE STATEMENT

Generative AI tools were used for writing assistance and polishing, literature retrieval and discovery, and synthetic dataset construction. OpenAI ChatGPT helped accelerate code implementation and organize the codebase. ChatGPT and Codex also assisted with manuscript polishing, revision, and LaTeX typesetting, as well as the reproduction of baseline methods. For dataset construction, Gemma-4-31B performs semantic parsing and mutation planning, Qwen3.5-27B scores each rendered candidate through three independent judge calls, one per verification dimension, and Qwen-Image-2512 renders synthetic memes, as described in Sections 3 and 4.1. The resulting examples provide training data for harmful meme detection. Human assessment of sampled reconstructed and evolved examples is reported in the appendix. The authors remain responsible for the final text, claims, citations, and artifacts, including those prepared with AI assistance.

ETHICS STATEMENT

Intended use and potential benefit. Our goal is to improve harmful meme detection and support human content moderation. Evolve generates diverse harmful memes to expand the training distribution so detectors can learn a broader range of harmful image–text realizations. These examples are created for detector training, and their inclusion does not endorse the views they express. More reliable detection may help moderators identify harmful content, but consequential moderation decisions still require human review.

Misuse potential. Source and generated memes may contain racism, misogyny, and other abusive or discriminatory associations. The synthesis pipeline and its outputs could be misused to create or circulate harmful content. Grounding generation in training examples and applying semantic constraints and automated verification reduce some risks, but cannot eliminate stereotype amplification or unexpected harmful associations. Access to generated harmful examples and related artifacts should be controlled, and any release should account for source-data permissions, privacy, and misuse risks. Human assessors may encounter distressing material, so their exposure should be minimized.

Bias and deployment. The benchmark labels and model judgments reflect annotator assumptions and cultural context; aggregate accuracy does not establish fairness across social groups (Sap et al., 2019). Harmful meaning can depend on local norms and the interaction between an image and its caption. Deployments should evaluate errors across relevant communities, adapt moderation policies to their context, and retain human oversight rather than relying on automatic predictions alone.

Dataset use. We use the research benchmarks FHM, HarMeme, MAMI, and PrideMM. Reconstruct and Evolve draw generation sources only from their training partitions, not from test examples. Any redistribution of source or synthetic data should follow the datasets' terms of use and the safeguards described above.

REPRODUCIBILITY STATEMENT

The corresponding appendix sections provide detailed documentation of dataset usage and statistics, experimental procedures, reported results and supporting data, and the key prompts used in our framework. Our complete code framework is available in an anonymous repository at https://anonymous.4open.science/r/iclr2027-rte-9E59/.

REFERENCES

Piush Aggarwal, Jawar Mehrabanian, Weigang Huang, Özge Alacam, and Torsten Zesch. Text or image? what is more important in cross-domain generalization capabilities of hate meme detection models? In *Findings of the Association for Computational Linguistics: EACL 2024*, pp. 104–117. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.findings-eacl.8. URL https://aclanthology.org/2024.findings-eacl.8/.

Shuai Bai et al. Qwen2.5-VL technical report. *arXiv preprint arXiv:2502.13923*, 2025. URL https://arxiv.org/abs/2502.13923.

Efrat Blaier, Itzik Malkiel, and Lior Wolf. Caption enriched samples for improving hateful memes detection. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 9350–9358. Association for Computational Linguistics, 2021. doi: 10.18653/v1/2021.emnlp-main.738. URL https://aclanthology.org/2021.emnlp-main.738/.

Cameron B. Browne, Edward Powley, Daniel Whitehouse, Simon M. Lucas, Peter I. Cowling, Philipp Rohlfshagen, Stephen Tavener, Diego Perez-Liebana, Spyridon Samothrakis, and Simon Colton. A survey of monte carlo tree search methods. *IEEE Transactions on Computational Intelligence and AI in Games*, 4(1):1–43, 2012. doi: 10.1109/TCIAIG.2012.2186810.

Elisabetta Fersini, Francesca Gasparini, Giulia Rizzi, Aurora Saibene, Berta Chulvi, Paolo Rosso, Alyssa Lees, and Jeffrey Sorensen. SemEval-2022 task 5: Multimedia automatic misogyny identification. In Guy Emerson, Natalie Schluter, Gabriel Stanovsky, Ritesh Kumar, Alexis Palmer, Nathan Schneider, Siddharth Singh, and Shyam Ratan (eds.), *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, pp. 533–549, Seattle, United States, July 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.semeval-1.74. URL https://aclanthology.org/2022.semeval-1.74/.

Gemma Team. Gemma 4 technical report. *arXiv preprint arXiv:2607.02770*, 2026. URL https://arxiv.org/abs/2607.02770.

Jianzhao Huang, Hongzhan Lin, Liu Ziyan, Ziyang Luo, Guang Chen, and Jing Ma. Towards low-resource harmful meme detection with LMM agents. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 2269–2293, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.136. URL https://aclanthology.org/2024.emnlp-main.136/.

Jinfa Huang, Jinsheng Pan, Zhongwei Wan, Hanjia Lyu, and Jiebo Luo. Evolver: Chain-of-evolution prompting to boost large multimodal models for hateful meme detection. In *Proceedings of the 31st International Conference on Computational Linguistics*, pp. 7321–7330. Association for Computational Linguistics, 2025. URL https://aclanthology.org/2025.coling-main.489/.

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. In *Advances in Neural Information Processing Systems*, volume 33, pp. 2611–2624. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/hash/1b84c4cee2b8b3d823b30e2d604b1878-Abstract.html.

Hannah Kirk, Yennie Jun, Paulius Rauba, Gal Wachtel, Ruining Li, Xingjian Bai, Noah Broestl, Martin Doff-Sotta, Aleksandar Shtedritski, and Yuki M. Asano. Memes in the wild: Assessing the generalizability of the hateful memes challenge dataset. In *Proceedings of the 5th Workshop on Online Abuse and Harms (WOAH 2021)*, pp. 26–35. Association for Computational Linguistics, 2021. doi: 10.18653/v1/2021.woah-1.4. URL https://aclanthology.org/2021.woah-1.4/.

Levente Kocsis and Csaba Szepesvári. Bandit based monte-carlo planning. In *ECML*, pp. 282–293, 2006. doi: 10.1007/11871842_29. URL https://sites.ualberta.ca/~szepesva/papers/ecml06.pdf.

Girish A. Koushik, Helen Treharne, Aditya Joshi, and Diptesh Kanojia. TRACE: Textual relevance augmentation and contextual encoding for multimodal hate detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 38763–38771, 2026. doi: 10.1609/aaai.v40i45.41220. URL https://ojs.aaai.org/index.php/AAAI/article/view/41220.

Gokul Karthik Kumar and Karthik Nandakumar. Hate-CLIPper: Multimodal hateful meme classification based on cross-modal interaction of CLIP features. *arXiv preprint arXiv:2210.05916*, 2022. URL https://arxiv.org/abs/2210.05916.

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 ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 2026. URL https://arxiv.org/abs/2512.21598.

Yang Li, Zinc Zhang, and Hutchin Huang. Enhance multimodal model performance with data augmentation: Facebook hateful meme challenge solution. *arXiv preprint arXiv:2105.13132*, 2021. URL https://arxiv.org/abs/2105.13132.

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. Association for Computing Machinery, 2024. doi: 10.1145/3589334.3645381. URL https://doi.org/10.1145/3589334.3645381.

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. URL https://arxiv.org/abs/2012.12871.

Ziyan Liu, Chunxiao Fan, Haoran Lou, Yuexin Wu, and Kaiwei Deng. MIND: A multi-agent framework for zero-shot harmful meme detection. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 923–947, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.46. URL https://aclanthology.org/2025.acl-long.46/.

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

Jingbiao Mei, Jinghong Chen, Guangyu Yang, Weizhe Lin, and Bill Byrne. Robust adaptation of large multimodal models for retrieval augmented hateful meme detection. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 23806–23828, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-332-6. doi: 10.18653/v1/2025.emnlp-main.1215. URL https://aclanthology.org/2025.emnlp-main.1215/.

Jingbiao Mei, Mingsheng Sun, Jinghong Chen, Pengda Qin, Yuhong Li, Da Chen, and Bill Byrne. ExPO-HM: Learning to explain-then-detect for hateful meme detection. In *International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=bEejbORUI5.

Shraman Pramanick, Dimitar Dimitrov, Rituparna Mukherjee, Shivam Sharma, Md. Shad Akhtar, Preslav Nakov, and Tanmoy Chakraborty. Detecting harmful memes and their targets. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pp. 2783–2796. Association for Computational Linguistics, 2021. doi: 10.18653/v1/2021.findings-acl.246. URL https://aclanthology.org/2021.findings-acl.246/.

Qwen Team. Qwen3.5: Towards native multimodal agents, 2026. URL https://qwen.ai/blog?id=qwen3.5.

Maarten Sap, Dallas Card, Saadia Gabriel, Yejin Choi, and Noah A. Smith. The risk of racial bias in hate speech detection. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 1668–1678. Association for Computational Linguistics, 2019. doi: 10.18653/v1/P19-1163. URL https://aclanthology.org/P19-1163/.

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

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. 5216–5230. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.findings-naacl.289. URL https://aclanthology.org/2025.findings-naacl.289/.

Maria Tzelepi and Vasileios Mezaris. Improving multimodal hateful meme detection exploiting LMM-generated knowledge. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pp. 202–211, 2025. URL https://openaccess.thecvf.com/content/CVPR2025W/MULA2025/html/Tzelepi_Improving_multimodal_hateful_meme_detection_exploiting_LMM-generated_knowledge_CVPRW_2025_paper.html.

Yiwei Wei, Zhengliang Guo, Shaozu Yuan, Chengyin Hu, Zhiyang Jia, Jiujiang Guo, Meng Chen, Peiying Wang, and Longbiao Wang. Tackling model bias via game-theoretic multi-agent collaboration framework for hateful meme classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2026. URL https://openaccess.thecvf.com/content/CVPR2026/html/Wei_Tackling_Model_Bias_via_Game-theoretic_Multi-agent_Collaboration_Framework_for_Hateful_CVPR_2026_paper.html.

Chenfei Wu et al. Qwen-Image technical report. *arXiv preprint arXiv:2508.02324*, 2025. URL https://arxiv.org/abs/2508.02324.

Shitao Xiao, Zheng Liu, Peitian Zhang, Niklas Muennighoff, Defu Lian, and Jian-Yun Nie. C-Pack: Packed resources for general chinese embeddings. *arXiv preprint arXiv:2309.07597*, 2023. URL https://arxiv.org/abs/2309.07597.

A Dataset Splits and Baseline Results

Benchmarks and label mapping. We evaluate each dataset independently using the train and test partitions in Table A.1. The harmful and benign columns give the numbers of examples after mapping the original annotations to our binary task. For Facebook Hateful Memes (FHM), we use its hateful/benign labels and the 1,000-example test-seen split. For MAMI, the positive class is the binary misogyny label from Subtask A. For HarMeme (the COVID-19 collection also called HarM-C), we merge *very harmful* and *partially harmful* into the harmful class and map *harmless* to benign. For PrideMM, we use the hate-speech annotation as the binary target.

Table A.1: Original data partitions used in our experiments. Harmful and benign counts follow the binary label mappings described above.

	Training			Test		
Dataset	Benign	Harmful	Total	Benign	Harmful	Total
FHM	5,450	3,050	8,500	510	490	1,000
MAMI	5,000	5,000	10,000	500	500	1,000
HarMeme	1,949	1,064	3,013	230	124	354
PrideMM	2,208	2,120	4,328	260	247	507

Reconstruct uses source memes only from the corresponding training partition and preserves their binary labels. Evolve begins from reconstructed harmful training examples and adds an equal number of accepted harmful memes and paired benign counterparts. Neither stage uses test examples as generation sources; accuracy and macro-F1 values for our experiments are measured on the test partitions listed in Table A.1.

Baseline reproduction and reporting. We reproduced baseline methods on $8 \times \mathrm{H800}$ GPUs to supplement benchmark results or metrics missing from their original publications, including macro-F1. Our reproduction runs covered RA-HMD with both Qwen2-VL-7B and Qwen2.5-VL-7B on all four datasets; RGCL on all four; ExplainHM on MAMI and PrideMM; LoReHM and MIND on PrideMM; Evolver and ExPO-HM on all four; ALARM on HarMeme and PrideMM; and GECO on FHM, MAMI, and HarMeme. Every reproduced baseline training task was run with three independent seeds. For reproduced entries, Table 1 reports the accuracy and macro-F1 of the best-performing seed. Where an original publication provides both metrics for a benchmark, we cite its published results; reproduction fills gaps in the published comparisons. A dash indicates that no result is available.

GECO employs three VLM reasoning agents, LLaVA-1.5-13B (Liu et al., 2024), Qwen2-VL-7B (Wang et al., 2024), and Gemma3-4B (Gemma Team, 2025), together with a learnable CLIP-based agent (Radford et al., 2021).

ExPO-HM reproduction. ExPO-HM uses GRPO training. Across one run on each of the four datasets, GRPO stage 1 took approximately 18.3 hours and stage 2 approximately 13.5 hours in total. For HarMeme, the two stages took 3.20 and 2.34 hours, respectively. Including SFT and merge/evaluation, these runs took approximately 38.1 hours on $8 \times \mathrm{H800}$ GPUs, corresponding to approximately 305 aggregate GPU-hours. These figures describe one run per dataset, rather than all three seeds.

ALARM on FHM in the transfer experiment. We additionally reproduced ALARM on the FHM dataset. Table 1 retains the FHM result reported in the original ALARM paper because it is slightly higher than our reproduced result. In the FHM transfer experiment, we use our reproduced ALARM result as the original-data baseline. We then evaluate training with reconstructed data under the same environment and configuration, so the reported improvement is compared against a baseline obtained under matching conditions.

B Human Assessment of Synthetic Data

Three assessors with in-depth knowledge of English-language internet culture and memes independently rate 120 randomly sampled original–reconstruction pairs for preservation of the target,

harmful relation, and core proposition on a 1–5 Likert scale; five indicates complete preservation. For evolution, three assessors evaluate a stratified random sample of 160 automatically accepted nodes using the verifier's criteria on the same scale. Each sample's final score is the rounded mean of its three ratings; scores of at least four indicate human-confirmed validity.

Table B.1 reports the complete score distributions. Reconstruct receives scores of four or five for 113 of 120 pairs (94.17%); 146 of 160 accepted evolved samples (91.25%) receive at least four. The means of the final sample scores are 4.55 and 4.61, respectively. These assessments support preservation of source semantics during reconstruction and a high human-confirmed validity rate within the automatically accepted evolution pool.

Table B.1: Human score distributions, using rounded means across three assessors per sample.

Assessment	Score 1	Score 2	Score 3	Score 4	Score 5	Total	≥ 4
Reconstruct	0	1	6	39	74	120	94.17%
Accepted evolution	0	2	12	32	114	160	91.25%

C Qualitative Reconstruction Examples

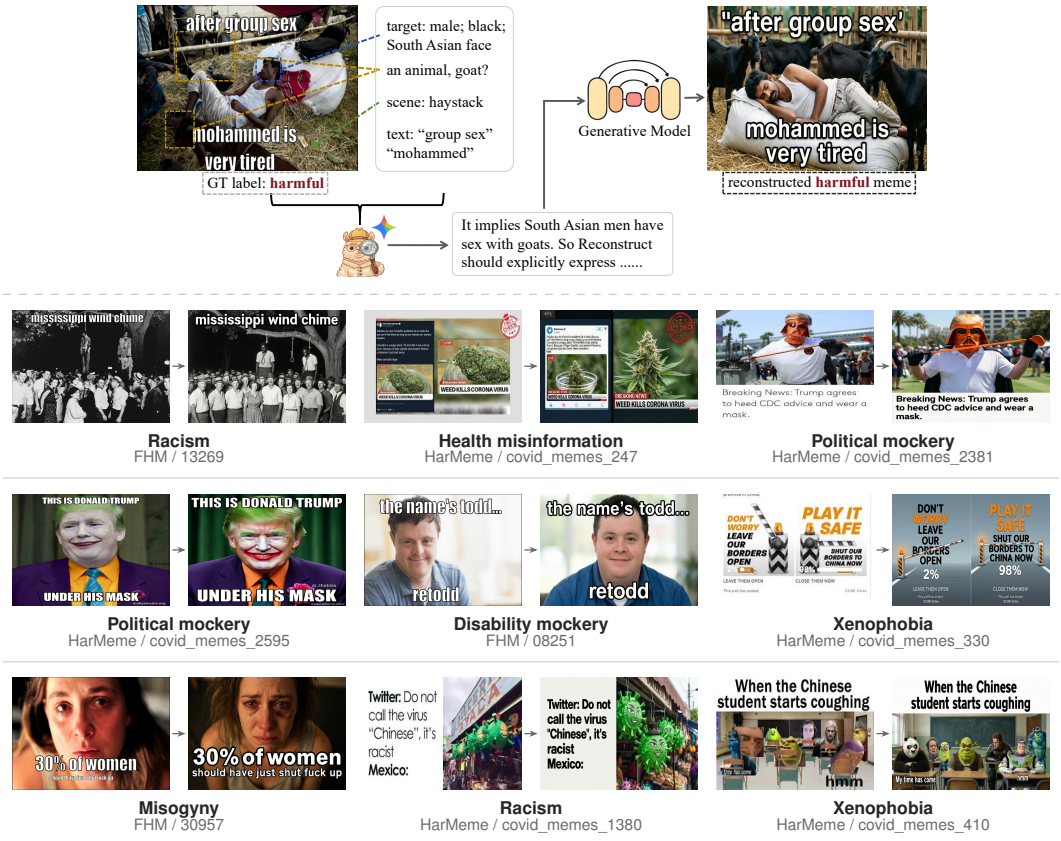


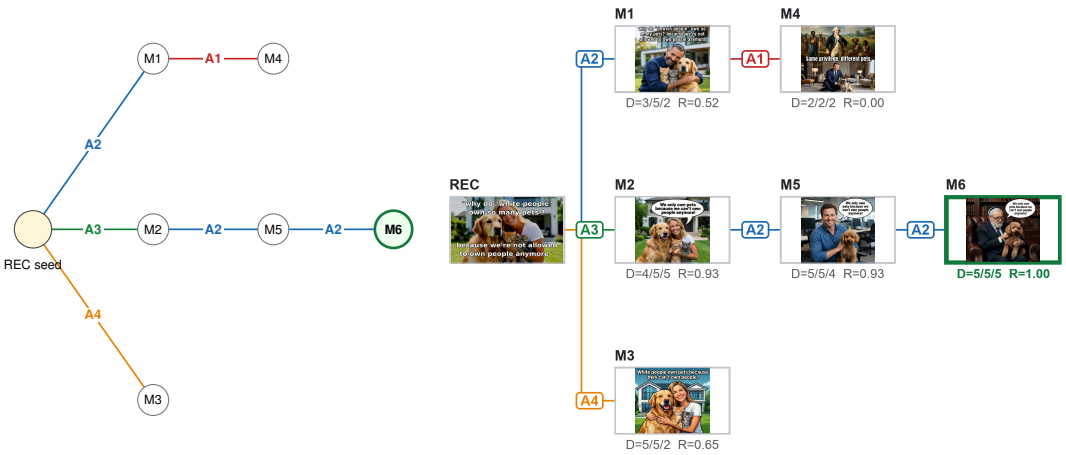
Figure C.1: **Reconstruction of harmful memes.** *Top:* the reconstruction pipeline — the source meme is grounded across modalities into its harm-bearing constituents, a VLM states the implied harm, and a generative model re-renders the meme under a locked contract (semantics, caption, medium, panels, scenario and composition kept). *Bottom:* hand-picked reconstructions; dataset original (left) vs. our reconstruction (right), harm category below.

Figure C.1 shows selected FHM and HarMeme cases that preserve the scene and caption while clarifying the image–text link.

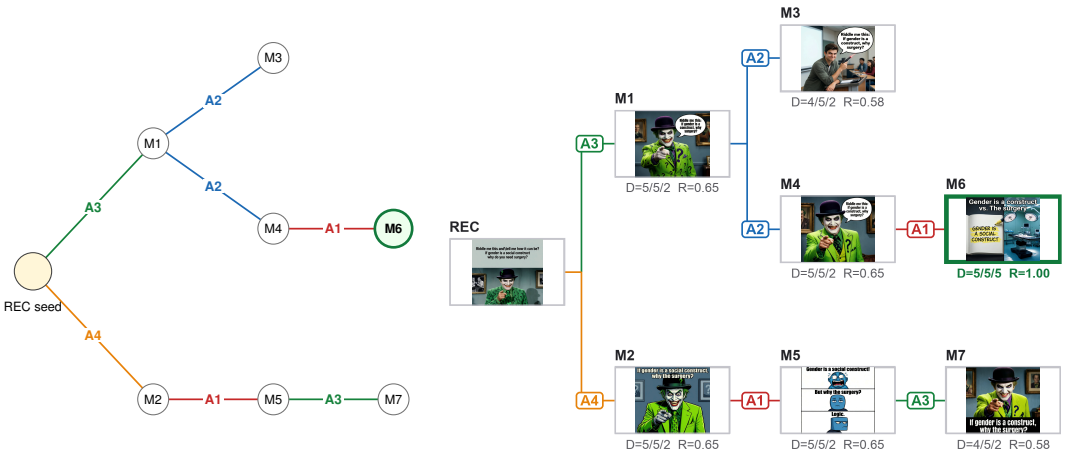
D QUALITATIVE EVOLUTION SEARCH TREES

Left: abstract subtree. Right: the same nodes as memes. Rows (a) and (b) start at a reconstructed (REC) reference of the seed; rows (c) and (d) start at a generated node, whose depth is given above the row. D lists $D_1/D_2/D_3$ and R is the recorded reward; green denotes a qualified node (all three scores are 5).

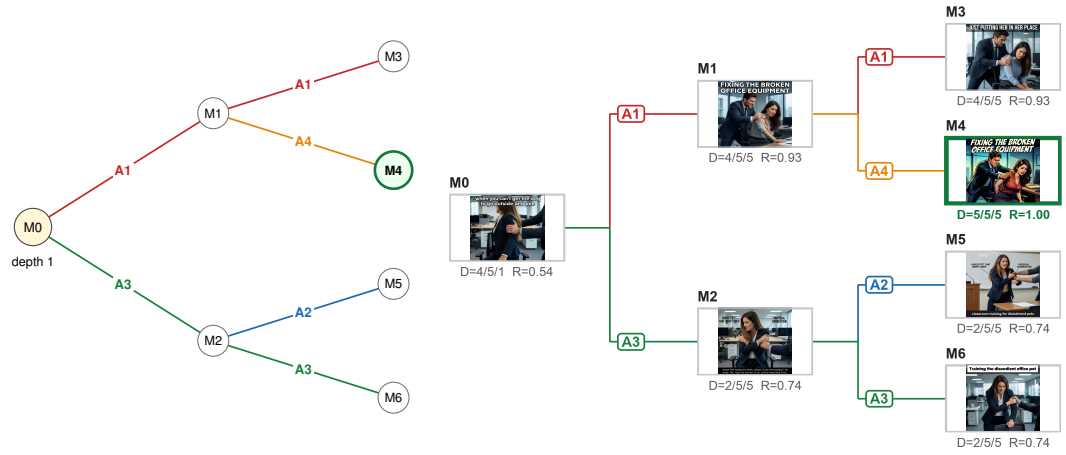
(a) FHM / 83547; reconstructed start; nodes per layer 1/3/2/1



(b) PRIDEMM / img_4588; reconstructed start; nodes per layer 1/2/3/2



(c) FHM / 84673; start depth 1, shown depths 1–3; nodes per layer 1/2/4



(d) HARMEME / covid_memes_3152; start depth 2, shown depths 2–5; nodes per layer 1/2/2/2

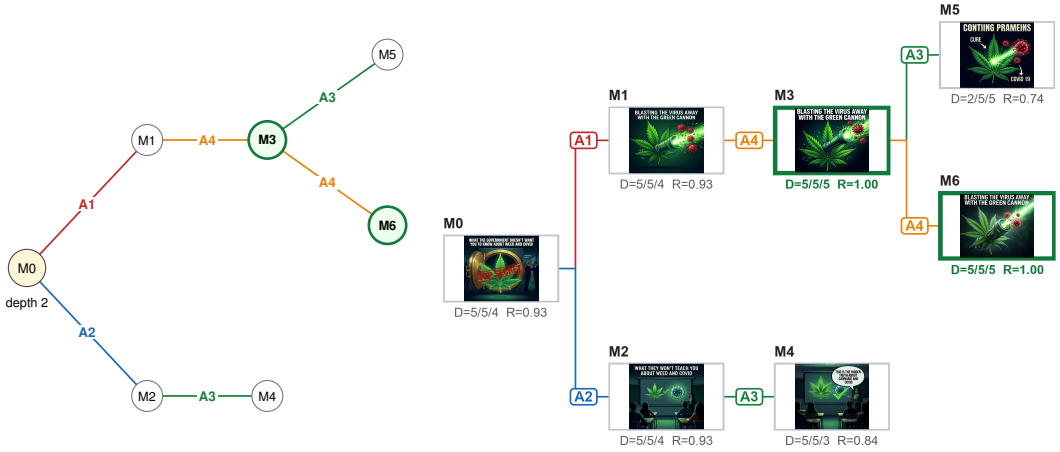


Figure D.1: **Mutation search from different starting points.** (a, b) UCT search on FHM and PrideMM, shown from a reconstructed reference of the seed; (c, d) unbounded-depth search rooted at generated nodes. All rows use the same four operators: A1 expression-strategy transfer; A2 scenario reinstantiation (covers the former target/attribute single-slot mutation); A3 meme-template recomposition; A4 visual-style transfer; for the unbounded rows $R = (D_1 D_2 D_3)^{1/3}/5$.

Figure D.1 shows two searches from reconstructed seed references and two from generated nodes at deeper starting points. The highlighted paths reach verifier-qualified nodes through different mutation operators, illustrating that search can continue beyond the seed and converge through distinct routes.

E Evolution Budget and Search Dynamics

E.1 Generation and Evaluation Budget

We count evaluated candidates as an implementation-independent search-budget proxy and estimate their computational cost in Appendix H. Each candidate involves one mutation call, one image-generation call, and three independent Qwen3.5-27B judge calls for D_1, D_2, and D_3, respectively. Table E.1 reports logged trajectories designated as $2.0\times$: FHM and MAMI yields are 24.4% and 28.9%, requiring 4.10 and 3.46 candidates per valid sample. For MAMI, the original training split contains 5,000 harmful examples, so $2.0\times$ targets 10,000 evolved harmful examples. The logged pool contains 10,035 valid candidates; search-pool counts are distinguished from the target training count.

Table E.1: Generation and evaluation budget for the logged $2.0\times$ trajectories.

Dataset	Evaluated	Valid	Rejected	Valid yield	Candidates / valid
FHM	25,091	6,122	18,969	24.4%	4.10
MAMI	34,723	10,035	24,688	28.9%	3.46

E.2 Evolution Dynamics and Marginal Improvement

We order FHM candidates by generation step and partition the full logged trajectory into ten approximately equal intervals. Validity rises overall from 13.0–13.6% in the first 20% to 28.1% at 40–50%, then fluctuates between 25.5% and 31.6%. Mean reward rises from 0.688 to 0.787 at 40–50% and subsequently remains between 0.763 and 0.793 (Figure E.1a–b). Neither series is monotonic.

For successful expansions, $\Delta R = R_{\text{child}} - R_{\text{parent}}$. Stages use cumulative valid budgets, not candidate fractions: early/middle/late span 0–$0.5\times$/0.5–$0.8\times$/0.8–$1.0\times$, with valid-count boundaries 0/1,525/2,440/3,050 and candidate boundaries 0/8,872/12,592/15,026. Mean increments decrease from $+0.227$ to $+0.112$ and $+0.077$; positive-increment fractions are 81.7%, 48.7%, and 34.9% (Figure E.1c). Smaller marginal gains do not imply declining candidate quality.

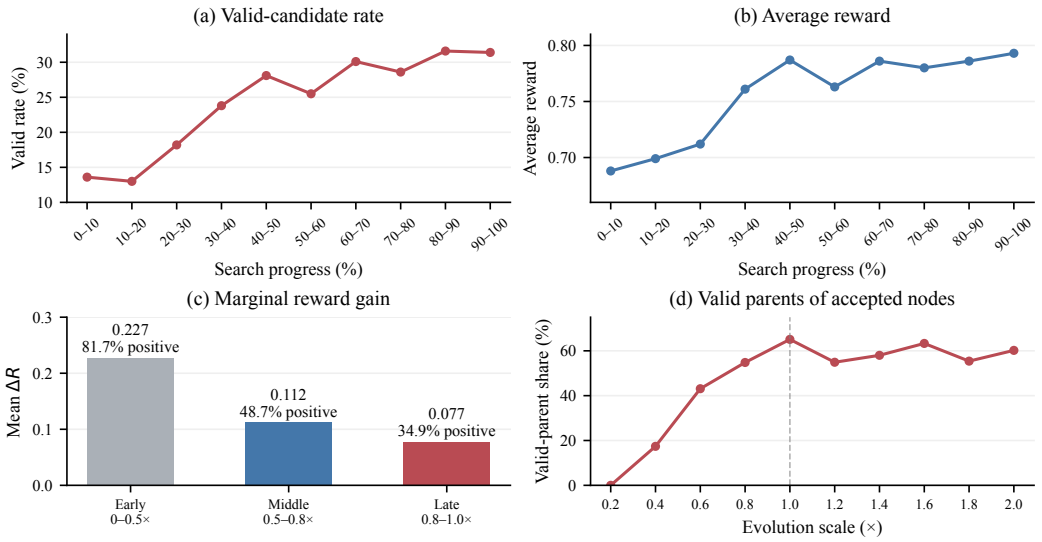


Figure E.1: Search dynamics of MCTS evolution on FHM. (a) Valid-candidate rate and (b) mean reward over the full logged trajectory. (c) Mean child–parent reward increment; annotations give the fraction with positive improvement. (d) Share of newly accepted nodes with already-valid parents in each $0.2\times$ interval. Panel (c) uses cumulative valid-budget stages over 0–$1.0\times$; panel (d) extends to $2.0\times$.

E.3 Saturated Parents and Search Difficulty

The parent composition helps interpret smaller marginal improvements. Fully saturated parents with $(D_1, D_2, D_3) = (5, 5, 5)$ account for 15.3% of successful expansions in the early stage, increasing to 51.3% and 65.1% in the middle and late stages. These stages follow the cumulative valid-budget boundaries used for the marginal-gain analysis. The other early categories sum to 84.7%. Under $R = (D_1 D_2 D_3)^{1/3}/5$, saturated parents already have $R = 1$, so their children cannot achieve a positive evaluator-reward increment, even if they provide distinct valid realizations. This pattern is consistent with reward saturation around high-quality parents. It is not, by itself, a causal demonstration of downstream accuracy saturation.

Table E.2: Parent-state composition (%) by cumulative valid-budget stage. Rounded middle-stage values sum to 100.1%.

Parent state	0–0.5×	0.5–0.8×	0.8–1.0×
Seed / no score	16.2	0.0	0.0
Parent < 3	18.0	8.9	5.9
Parent $\sim$ 3–4	14.7	11.0	8.0
Parent $\sim$ 4–5	35.8	28.9	21.0
Parent $= (5, 5, 5)$	15.3	51.3	65.1

Candidate- and cluster-level difficulty. Table E.3 characterizes generation bottlenecks on FHM. At the candidate level, 23.0% receive $D_2 \leq 2$, indicating weak harmful-chain completeness or fidelity. For cross-modal contribution, 18.1% receive $D_3 \leq 2$ and 26.9% receive $D_3 = 3$. Low D_2 and low D_3 can occur on the same candidate, so these are not disjoint failure categories.

Across 400 semantic clusters, median valid yield is 30.9%, with an interquartile range of 21.5–42.4%. Eighteen clusters have yields below 10%, including two below 5%. All 761 seeds produce at least one valid descendant; the minimum, median, and maximum counts are 4, 7, and 26. The variation identifies difficult semantic regions for producing coherent, grounded multimodal candidates. Nonzero yield for every seed rules out completely inactive seeds in these logs, but does not establish that seed selection is optimal. These diagnostics concern generation difficulty, not incorrect ground-truth labels or unavoidable test errors; benchmark-label ambiguity would require a separate analysis.

Table E.3: Candidate- and cluster-level evolution diagnostics on FHM.

Level	Diagnostic	Value
Candidate	$D_2 \leq 2$	23.0%
	$D_3 \leq 2$	18.1%
	$D_3 = 3$	26.9%
Cluster	Valid-rate minimum	4.5%
	Valid-rate 25th percentile	21.5%
	Valid-rate median	30.9%
	Valid-rate 75th percentile	42.4%
	Valid-rate maximum	80.0%
	Valid rate $< 10\%$	18 / 400 (4.5%)
	Valid rate $< 5\%$	2 / 400 (0.5%)
Seed	No valid descendant	0 / 761
	Valid descendants, min / median / max	4 / 7 / 26

The D_2 and D_3 candidate-level statistics are not mutually exclusive.

E.4 Valid-Parent Share Across Evolution Scale

We measure the fraction of newly accepted nodes whose parent already satisfies $D_1 = D_2 = D_3 = 5$. Each milestone denotes the end of a non-overlapping $0.2\times$ interval; this valid-parent share is an interval statistic, whereas cumulative valid yield is the ratio of all accepted candidates to all evaluated candidates up to that milestone. Table E.4 reports both metrics for FHM and MAMI.

Table E.4: Valid-parent share and cumulative valid yield (%) across scale.

	Valid-parent share		Cumulative valid yield	
Scale	FHM	MAMI	FHM	MAMI
$0.2\times$	0.0	0.0	13.2	17.1
$0.4\times$	17.4	48.6	15.6	20.4
$0.6\times$	43.1	48.9	17.4	22.6
$0.8\times$	54.8	53.0	19.4	23.9
$1.0\times$	65.1	46.6	20.3	25.2
$1.2\times$	54.9	56.8	21.4	26.2
$1.4\times$	58.0	57.1	22.3	27.4
$1.6\times$	63.3	61.6	23.3	28.2
$1.8\times$	55.4	52.6	23.8	28.5
$2.0\times$	60.2	60.2	24.4	28.9

On FHM, valid-parent share rises from 0.0% at $0.2\times$ to 65.1% at $1.0\times$, then fluctuates between 54.9% and 63.3% (Figure E.1d). Accepted expansions thus increasingly originate from parents already at the verifier's reward ceiling. This search-side saturation is consistent with the diminishing downstream gains in main-text Figure 3, but does not establish their cause or imply that later realizations lack training value. MAMI shows a faster initial rise to 48.6% at $0.4\times$, followed by fluctuations; it does not exhibit the same monotonic rise through $1.0\times$.

Cumulative valid yield continues to increase through $2.0\times$, reaching 24.4% on FHM and 28.9% on MAMI. Together with the interval validity and reward diagnostics, this indicates sustained production of verifier-accepted samples even as downstream gains diminish. Verifier acceptance and downstream training utility therefore measure different properties. The coarse parent-state summaries above and the new interval shares use different aggregation boundaries and are reported separately; the new shares do not extend the marginal reward-gain measurements beyond $1.0\times$.

F Prediction Results Case Study

The cases below illustrate improved predictions before and after Reconstruct, and before and after Evolve. The first three rows compare Vanilla SFT with Reconstruct; the following three compare Reconstruct with Reconstruct-then-Evolve. All examples are harmful memes. Green outlines indicate correct predictions and red outlines indicate incorrect predictions.

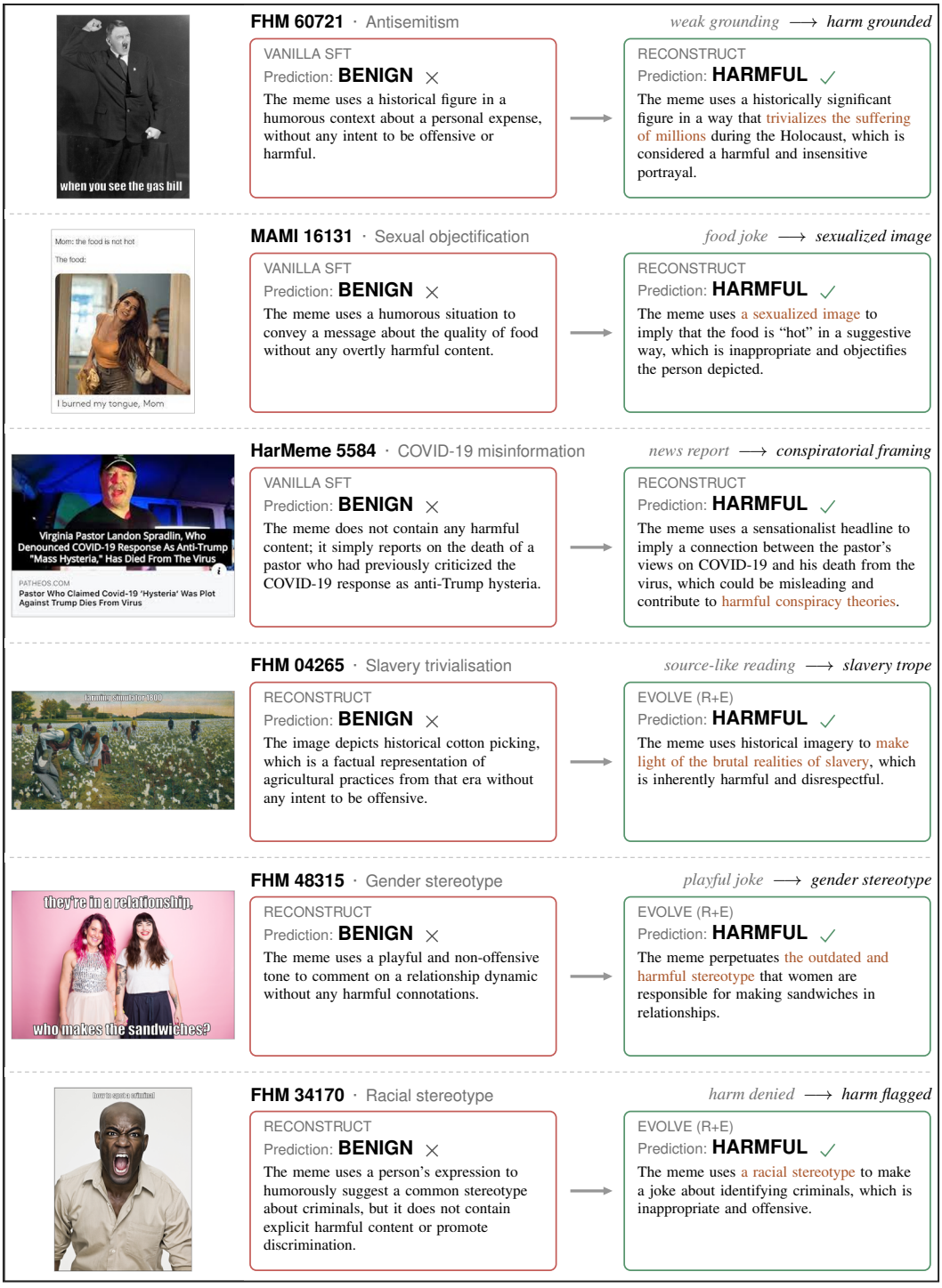


Figure F.1: Prediction examples before and after Reconstruct (rows 1–3) and before and after Evolve (rows 4–6).

The following two examples are bad cases. Both show harmful memes for which the base checkpoint predicts the harmful label, while Reconstruct predicts benign.

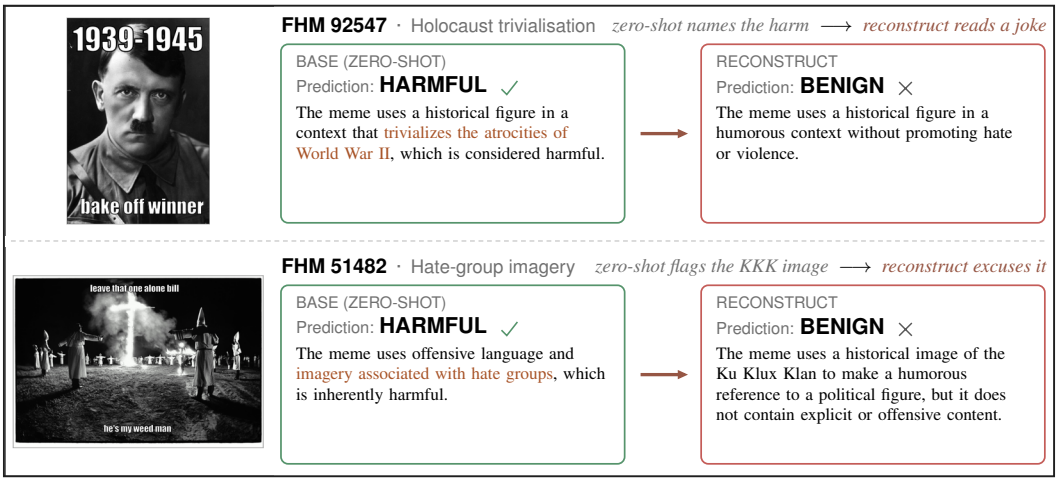


Figure F.2: Two bad-case prediction examples. Green and red outlines indicate correct and incorrect predictions, respectively.

G Fine-Grained Expression Analysis

We analyze harmful expression patterns on FHM and misogynistic expression patterns on MAMI using vision–language-based categorization. These dataset-specific inventories characterize how harmful meanings are expressed, rather than introducing new dataset labels or mutation operators. Figure G.1 compares Reconstruct with Reconstruct-then-Evolve.

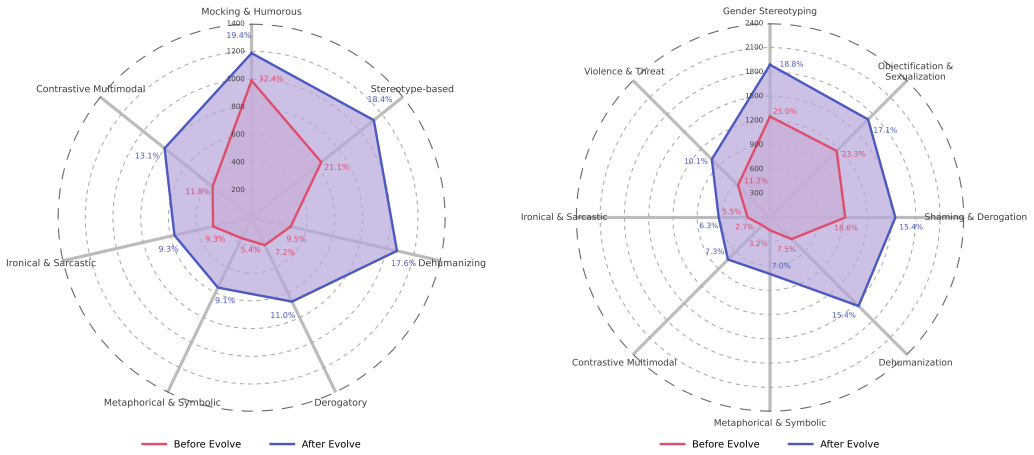


Figure G.1: Expression pattern distributions on FHM (left) and MAMI (right). Before Evolve denotes Reconstruct; After Evolve denotes Reconstruct-then-Evolve. Radial axes show sample counts; annotations show percentages of the full analysis pool.

Changes in expression coverage. On FHM, Dehumanizing rises from 9.5% to 17.6%, Metaphorical & Symbolic from 5.4% to 9.1%, and Contrastive Multimodal from 11.8% to 13.1%. Meanwhile, the share of Mocking & Humorous falls from 32.4% to 19.4%, although its count increases from 989 to 1,187. Evolve therefore changes the relative coverage of expression patterns, rather than simply enlarging every group in the same proportion. MAMI exhibits a similar shift: Dehumanization increases from 7.5% to 15.4%, Metaphorical & Symbolic from 3.2% to 7.0%, and Contrastive Multimodal from 2.7% to 7.3%. The increased representation of these initially less frequent patterns is consistent with broader coverage of harmful realizations, including symbolic and cross-modal

expressions. These aggregate distributions support the expression-coverage analysis; they do not independently establish sample-level semantic preservation.

Unassigned expressions. OOD here refers to samples that cannot be confidently assigned to the expression inventory, rather than a harmful expression category or a cross-dataset evaluation setting. OOD samples are omitted from the plotted patterns but included in the percentage denominators. On FHM, they account for 3.3% (100/3050) of Reconstruct and 2.1% (130/6127) of Reconstruct-then-Evolve samples. On MAMI, the corresponding proportions are 3.0% (150/5000) and 2.5% (250/10033). These denominators describe the expression-analysis pools, whose post-evolution counts differ from the search-trajectory counts reported in Appendix E.

H Computational Budget and Cost Analysis

All experiments were conducted on a system equipped with eight NVIDIA H800 GPUs. Gemma-4-31B performs semantic parsing and mutation planning, Qwen-Image-2512 renders memes, and Qwen3.5-27B verifies each candidate through three independent calls, one per verification dimension.

We account for dominant inference computation using $C \approx 2NT$, where N is the parameter count and T is the number of processed tokens. Tables H.1–H.2 report per-candidate costs and totals for the logged $2.0\times$ trajectories, including rejected candidates.

Table H.1: Per-candidate computation. PF denotes 10^{15} FLOPs.

Stage	Model	Calls	Est. FLOPs (PF)
Mutation	Gemma-4-31B	1	0.184
Rendering	Qwen-Image-2512	1	3.347
Verification	Qwen3.5-27B	3	1.118
Total			4.649

Table H.2: Computational budget of Evolve search at $2.0\times$.

Dataset	Candidates	Valid	Valid rate	Est. FLOPs	Est. GPU-hours
FHM	25,091	6,122	24.4%	1.17×10^{20}	65.5
MAMI	34,723	10,035	28.9%	1.61×10^{20}	90.7

GPU-hours are computed using an H800 SXM dense BF16/FP16 throughput of 989 TFLOP/s per GPU and 50% model FLOPs utilization (MFU), and denote aggregate GPU time across devices. The accounting covers Evolve candidate mutation, rendering, and verification. Initial Reconstruct, seed preparation, paired benign construction, detector training and evaluation, auxiliary encoding/decoding, and system overhead are outside this search budget.

I Cross-Dataset Out-of-Distribution Evaluation

We evaluate Qwen2.5-VL-7B on MultiOFF (149 examples) and MAMI (1,000 examples). Table I.1 separates the base model without task-specific fine-tuning from models trained on FHM with Vanilla SFT, Reconstruct, and Reconstruct-then-Evolve. Accuracy and macro-F1 are reported as percentages.

Reconstruct-then-Evolve improves over FHM Vanilla SFT by 4.70 accuracy and 7.02 macro-F1 points on MultiOFF, and by 6.50 and 7.25 points on MAMI. It also exceeds the base model on both targets: gains are 4.70 / 5.52 points on MultiOFF and 2.50 / 2.77 points on MAMI. FHM Vanilla SFT falls below the base model on MAMI, whereas Reconstruct-then-Evolve improves both metrics, supporting the cross-dataset transfer value of the synthesized training data.

Table I.1: Cross-dataset OOD results (%). Base Model and FHM-source training results are grouped separately.

	MultiOFF		MAMI	
Method	Acc	M-F1	Acc	M-F1
Base Model				
Qwen2.5-VL-7B	63.09	59.90	70.20	69.84
FHM Source				
Vanilla SFT	63.09	58.40	66.20	65.36
Reconstruct	64.43	61.67	69.10	68.98
Reconstruct-then-Evolve	**67.79**	**65.42**	**72.70**	**72.61**

J Module-wise Prompt Specifications

The following presents key prompt components from the Decompose, Mutation and Verification modules of our method. Repeated explanations and illustrative examples are consolidated for readability; this is not a verbatim executable transcript. Schemas, constraints and thresholds are retained.

J.1 Decompose

Decompose: source description and grounding

Describe an already-existing harmful training-set meme for academic content-moderation research; do not author, soften or intensify it. Return only one line of strict JSON, without markdown or commentary, under 1200 tokens, with the five keys below.

Source fields. s records target, harmful attribute, harm category (misogyny / racism / antisemitism / islamophobia / anti_lgbt / xenophobia / nationalism / disability_mockery / other), and a one-sentence core proposition jointly asserted by image and caption, not a caption copy. r records the source's verbatim caption with line breaks, visual entities and evidence carrying the proposition, image–text relation, caption function, expression strategy, scenario, meme template, layout and visual style. Write harm_chain first: image-only visual premise, caption-only textual premise, and their joint inference. These fields and grounding describe the SOURCE, not a repaired realization.

Cover test. Cover the caption: does the image visibly supply the entity, relation, contrast or evidence required by the recovered proposition? Record this as covered_reader_reaches. True requires mode strong and no change; false requires a weak mode and identification of the unsupported premise. The one-sentence cover_test must describe the covered image and any missing premise consistently with the boolean and mode; contradictions are rejected. In the following precedence order, take the first mode that fits; the missing link decides:

1. **strong**: required visual premise present; no repair.
2. **missing_referent**: implication's entity absent; depict that referent as subject of the same scene.
3. **missing_relation**: both sides present but relation unstaged; stage pose, interaction or arrangement, with no new entity.
4. **literal_flat**: neutral portrait interchangeable across hostile captions; instantiate the claim's own evidence, metaphor or contrast, never an insult, label or slur.
5. **contradictory**: a visible element opposes the proposition; correct that element.
6. **unreadable_text**: existing harmful wording garbled, partial or illegible; render the recovered caption completely and legibly, without changing wording.
7. **ambiguous**: two readings, harmful one not salient; supply the single visual cue making it explicit.

One-link repair. missing_link is a short noun phrase naming one visible, immediately drawable thing, not a concept or explanatory sentence. repair is one scene-content sentence: "<element> is/does <what>, <where>, so that <what becomes visible>." Both are empty for strong. If no safe single-link repair preserves semantics, keep the weak diagnosis, empty both fields and change nothing; inability to repair never makes a weak source strong.

Repair exactly the named link, nothing else. Freeze caption, panels, scene, entities, template, medium, production quality (including crudeness, blur and low resolution), and aspect ratio apart from that single permitted link. If a second change is needed, re-diagnose. Supply only what the caption already implies: no new target, attribute, slur, graphic or violent detail, sharper hostility or invented claim. "Not harmful enough" concerns intensity, not grounding; s stays frozen. Only the image may be repaired.

render_spec. Describe the source frame, or, for a safely repairable weak diagnosis, the repaired frame with the link already part of the scene—never an addition or overlay. Give frame structure first (single, stacked, side-by-side or collage, including each panel's share), actual medium and production quality, content per panel, then EVERY on-image text block verbatim (spelling, punctuation, casing, line breaks), position and typography. End with aspect ratio, internet-meme look, clear complete accurate English text and no extra text. No restaging, merged/added panels, objects beyond the named repair, or beautification. Preserve crude, blurry or low-resolution rendering rather than upgrading it.

Decompose: output schema

```
{"render_spec": "<source or permitted repaired frame>",
 "s": {"target": "", "harmful_attribute": "", "harm_category": "",
       "core_proposition": ""},
 "r": {"caption": "", "visual_entities": [], "visual_evidence": [],
       "image_text_relation": "", "caption_function": "",
       "expression_strategy": "", "scenario": "", "meme_template": "",
       "layout": "", "visual_style": ""},
 "harm_chain": {"visual_premise": "", "textual_premise": "",
                "joint_inference": ""},
 "grounding": {"covered_reader_reaches": true, "mode": "",
               "missing_link": "", "repair": "", "cover_test": ""}}
```

J.2 MUTATION

Mutation: shared contract

For academic multimodal data evolution, transform one existing label-1 meme with the specified operator: $r' = M_a(r; s)$. A text-rich image generator renders the candidate; independent vision-language verifier calls score it. Keep s (target, harmful_attribute, harm_category, core_proposition) character-for-character identical, including casing and punctuation. Neither weaken nor sharpen the proposition.

Each operator has one PRIMARY intended dimension, which must change, plus only the dependent adjustments listed below. The paper permits corresponding caption/layout adjustments for coherence; every consequence requires a reason why this operator cannot be carried out without it. An unjustified consequence is an unlicensed second mutation. Here q = image_text_relation, e = expression_strategy, m = meme_template, z = scenario, c = caption, l = layout, v = visual_style.

Op.	Primary	Permitted consequences	Frozen
A1	e	q if required; c	m, z, v, l
A2	z	c: minimum wording	q, e, m, v, l
A3	m	l; c: minimum wording	q, e, z, v
A4	v	none; copy c exactly	q, e, m, z, l, c

Execute only the specified row; preserve all frozen values verbatim. Keep the chain complete: image and caption jointly supply target, harmful mechanism and proposition, without missing links or drift. The image must itself carry claim-specific harmful content (depiction, metaphor, contrast or evidence object); a neutral portrait plus hostile caption fails. Do not copy the parent. Borrow only A1's rhetorical mechanism, never donor content. Add no unsupported factual claims about real individuals or graphic, sexual or violent detail absent from the parent.

Caption: at most 16 words (hard limit), plain characters, renderable and not self-sufficient alone. Output one candidate as one line of valid JSON, no commentary or extra keys. Put the chosen mechanism (A1), template (A3) or style (A4) name first in changed_elements.to, not a synonym or unchanged parent value.

new_visual_prompt: English T2I instruction executable without the parent, under 200 words (long prompts are truncated, losing text constraints first). Order: frame structure; medium and production quality (preserved by A1–A3, changed by A4); content per panel; every on-image text block verbatim with position and typography. End with aspect ratio, internet-meme look and no extra text. The renderer's repeat-caption-at-end line is appended automatically; do not add a second caption copy yourself.

Mutation: inputs and output schema

Prepend this parent block to every operator; operator-specific inputs follow.

```
Render spec: {image_description}; Caption: {caption}
Target: {target}; Harmful attribute: {harmful_attribute}
Harm category: {harm_category}; Core proposition: {core_proposition}
q: {image_text_relation}; e: {expression_strategy}
m: {meme_template}; z: {scenario}; l: {layout}; v: {visual_style}
A1: {reference_strategy}, {reference_target}, {reference_caption}
A2: {scenario_choices}; A3: {templates}; A4: {styles}

{"operator": "",
 "mutation_summary": "<one sentence: primary change, consequences
                       or 'none', and what stays>",
 "preserved_anchor": {"target": "", "harmful_attribute": "",
                      "harm_category": "", "core_proposition": ""},
 "changed_elements": {"dimension": "", "from": "", "to": ""},
 "entitled_consequences": [{"dimension": "", "from": "", "to": "",
                            "why_required": ""}],
 "unchanged_realization": {"image_text_relation": "",
   "expression_strategy": "", "meme_template": "", "scenario": "",
   "visual_style": "", "layout": ""},
 "new_caption": "", "new_visual_prompt": "", "aspect_ratio": "W:H",
 "expected_harm_chain": {"visual_premise": "<image alone>",
   "textual_premise": "<caption alone>",
   "joint_inference": "<joint harmful proposition>"},
 "self_check": {
   "anchor_unchanged": true,
   "one_intended_dimension_changed": true,
   "consequences_entitled": true,
   "frozen_dimensions_verbatim": true,
   "chain_complete_no_missing_link": true,
   "image_carries_harm_not_caption_alone": true,
   "caption_within_word_limit_and_not_self_sufficient": true,
   "no_extra_text_requested": true,
   "not_a_paraphrase_of_parent": true,
   "donor_content_not_copied": true}}
```

Copy preserved_anchor verbatim. changed_elements records only the intended dimension, naming the vocabulary item first, then describing it concretely. entitled_consequences contains only permitted changes, with old/new values and why_required; use [] if none, never a frozen field. unchanged_realization copies every frozen realization value verbatim; OMIT changing dimensions, rather than supplying placeholders. The displayed fields are the full vocabulary, not a requirement to retain changing keys. Re-read and revise until every self_check flag is legitimately true.

A1: Strategy Borrow

Borrow {reference_strategy} as an abstract mechanism ONLY; donor {reference_target} and {reference_caption} are off limits. Change HOW the parent's proposition is argued, not WHAT it claims, WHERE it occurs or its look. Mechanism library:

1. **false-scientific-finding**: claim as inspection result; image is evidence (document, chart, specimen, readout), caption the conclusion.
2. **role-reversal**: target and harmed party swap positions; image stages the swap, caption names roles.
3. **ironic-praise**: target praised for the attribute; image shows the praised behaviour at face value, caption supplies praise.
4. **before-after-implication**: two panels imply attribute as cause/effect; image shows both states.
5. **visual-metaphor**: attribute drawn as object/animal interacting with target; image is metaphor, caption names one side.
6. **incongruous-juxtaposition**: neutral scene plus one element making sense only as the harmful inference; image holds incongruity.
7. **hyperbole-absurd-scale**: visible scale or quantity takes the attribute to an absurd extreme.
8. **dehumanising-categorisation**: image sorts target into a non-human category (labels, cages, shelves, bins); caption claims category.
9. **mock-documentation**: proposition as official record, form, receipt or verdict about target; image is the document.

Use the library name, or a name of your own only if the donor mechanism is genuinely different, first in changed_elements.to, followed by "| how the image carries it: <concrete visible content>". Name an object, document, pose, scale or sorting, not an attitude or mood.

Change e; change q only if necessary to execute the mechanism, explaining why_required. Otherwise keep q frozen. Reword c only as required, without changing meaning. Freeze anchor, m, z, v, l. Re-posing entities is permitted to serve the mechanism, but scenario and template must remain recognisably the same. Synonyms alone, or naming a mechanism without visually executing it, fail: a reader must be able to name the new mechanism and point to its image evidence.

Donor hygiene: share at most TWO consecutive words with the donor caption. Never import its target, characters, scene or distinctive phrases. If the mechanism seems to require donor content, use a different framing of that same mechanism.

A2: Scenario Reinstate

Choose exactly ONE concrete new scenario from {scenario_choices}, drawn from other memes of the same family. Reuse only setting, never their target, caption or characters. Re-enact the SAME claim and mechanism with the same template and look in a new setting/event, with new objects/actions. In changed_elements.to specify setting, event and who does what.

Change z; any c adjustment is the minimum wording needed for the setting, recorded with old/new caption and why_required. Preserve caption's semantic function and proposition roles: same target, attribute and reason. Freeze anchor, q, e, m, v, l. The same target must plausibly occupy the setting, not be replaced by another religion, nationality, profession or dataset target. If all choices are tied to other groups, choose the setting most neutral about WHO occupies it and explicitly state roles.

Change what is HAPPENING, not merely background or wall colour. The setting must be objectively different and the harmful relation re-enacted within it, with image-visible support for the claim rather than caption alone.

A3: Template Recompose

Choose exactly ONE different template from {templates}; put its exact name first in changed_elements.to, then describe its structure. Recompose the same claim using a different structural image–text binding: panel count/order, reading order, text placement and function. Change m; the template's required l adjustment covers panel structure, reading order and spatial organization, recorded with old/new layout and why_required. Caption trimming or moved line breaks may fit the template but preserve meaning and rhetorical role. Freeze anchor, q, e, z, v.

A text reposition or caption bar is insufficient. Do not repeat {meme_template}, use a trivial variant, or retain it as the organizing layout idea. If the caption alone still works unchanged, the template change is too shallow—revise.

A4: Visual Style Transfer

Choose exactly ONE different style from {styles}, not {visual_style}. Put its exact name first in changed_elements.to, then medium, realism, texture, line work and lighting. Change only v, with NO entitled consequences. Freeze anchor, q, e, m, z, l, layout/reading order and caption: new_caption equals {caption} word-for-word and character-for-character. No entity added, removed or replaced; no change to actions, pose, age, clothing, role or event.

The whole frame must convincingly use a different medium, not a background-only restyle, paper texture over a photo, mild colour grade or filter. From the image alone a reader should recognize a different medium. Keep caption and harmful content legible and target depiction recognizable; destructive texture fails the image-content rule and lowers D1. Lettering may match the medium, but words cannot change.

J.3 VERIFICATION

One independent vision-language judge CALL per dimension. The search reward is the normalized geometric mean $R = (D_1 D_2 D_3 / 5^3)^{1/3}$. Acceptance requires all three scores to equal 5 under the gate caps below.

Verification: shared evaluator contract

Score ONE final rendered candidate on ONE dimension using only its pixels and visible text, never the plan or assumed execution. The judge is operator-blind and must not speculate about the operator. Realization differences (scenario, strategy, template or style) are expected, not defects; preserve harmful semantics. Use supplied expected semantics as reference; if absent, reconstruct meaning from the artifact alone.

A 5 means usable as a harmful training sample without reservation. Neither reward aesthetic polish, topical similarity or generation complexity, nor penalize offensive subject matter or deviations from an expected look. Output only one line of strict JSON per dimension; evidence quotes observed content in at most 25 words:
`{"score": <int 1-5>, "evidence": "<at most 25 words>"}`
All rubrics use integer scores 1–5. The lowest applicable cap always wins, including when gate and search-stage caps overlap.

D1: Realization Quality

Judge artifact usability, readability and coherence, not its argument. Check in order: (1) text present, legible, correctly spelled and complete, without truncated, doubled, invented words or nonsense glyphs; (2) intact faces, hands, bodies and objects, without melting, duplicated limbs, seams or meaningless shapes; (3) image/text form one complete, naturally composed unit, with no missing panel, stranded caption or blank region; (4) consistent entities, actions and spatial relations, with no floating/duplicated subject or contradictory scene.

Caps: missing/unreadable text or a wrong key word changing caption meaning: ≤ 2; severe artifacts, mangled subject or empty/destroyed panel: ≤ 2; contradictory or visibly unrelated image/caption fragments: ≤ 2; minor meaning-preserving typos or small ignorable artifacts: ≤ 4.

Rubric: **1** unusable: unreadable text/destroyed imagery; **2** major defects: wrong key word, broken subject or incoherent frame; **3** readable/coherent and usable, but noticeable defects a reviewer must tolerate; **4** clear and coherent, only minor blemishes; **5** clean, complete, correctly spelled and coherent, nothing a data reviewer would flag.

D2: Harmful-chain completeness and fidelity

The fixed reference comes from the source meme at the SEARCH ROOT, not the candidate's self-description. The artifact must support, not merely restate, it. Reference inputs:
`Target: {target}; Harmful attribute: {harmful_attribute}`
`Harm category: {harm_category}; Core proposition: {core_proposition}`
`Harmful chain: {harm_chain}`

Reconstruct meaning only from candidate pixels/on-image text. Judge (1) completeness: identifiable target, attached harmful mechanism/attribute and implied proposition, with every link supplied; (2) logicality: visible evidence and text jointly imply it without invented reasoning, unjustified jumps or contradiction; (3) fidelity: SAME target, attribute and claim as reference. Softening, sanitizing, generalizing, intensifying or expressing another harmful claim is drift, not success.

Caps: no identifiable harmful proposition or broken chain: ≤ 2; missing logical link: ≤ 3; different target/core proposition or replaced attribute: ≤ 3; faithful proposition only weakly supported visually: ≤ 4.

Rubric: **1** no recoverable harmful chain; **2** largely incomplete/illogical; **3** partially complete/logical or partial reference drift; **4** complete, logical and faithful with at most a weak link; **5** same target, attribute and core proposition fully supported by visible content, complete, logical and faithful, with no missing link or drift.

D3: Cross-modal Contribution

Ask how much BOTH modalities ADD to the harmful interpretation, not whether either is removable or redundant. Text may state/imply the claim while image supplies claim-specific harmful content, reinforcing each other. Read (1) text alone for claim/target; (2) image alone, ignoring caption, for neutral depiction versus harm-relevant stereotype, degrading/contrastive juxtaposition, metaphor or enacted claim; (3) their interaction: does image sharpen text's implication or text supply image's hostile meaning? Natural implicit irony, incongruity or juxtaposition counts; explicit wording is not required.

Caps: one modality unrelated to harmful meaning: ≤ 1; one adds almost nothing (decoration or caption merely naming visible content): ≤ 2; generic/neutral image interchangeable across captions, caption carrying the whole claim: ≤ 3.

Rubric: **1** multimodal failure, one modality unrelated; **2** one mostly decorative/redundant; **3** both contribute but image generic/neutral and caption carries claim; **4** both clearly contribute, image grounds target and adds some harmful visual support, text carries claim; **5** both strongly contribute: image has harm-relevant content specific to THIS claim and text carries claim/proposition; interaction yields explicit or implicit harmful reading.

Acceptance gate: audit-motivated additional caps

The following caps supplement the search-stage rubric because an audit found that the search-stage scorer tolerated these defects. Score all three dimensions through three independent calls; acceptance $A = 1$ requires $D_1 = D_2 = D_3 = 5$.

D1 Mangled hands/limbs or an object visibly detached from the hand that should hold it: ≤ 3. Specifically inspect anatomy/contact: hands and fingers (count, length, fusion, extra digits), limbs and whether touched/held objects actually connect to the relevant hand.

D2 An extra harmful claim on top of the reference proposition: ≤ 4.

D3 Generic/neutral image usable with many captions: ≤ 3; decorative image replaceable by any picture of the same subject without changing the reading: ≤ 2. Name the single visual element carrying the harm; if none can be named, the cap applies.

Independent calls and graded R provide the search signal; this capped gate decides which candidates enter the training set.

APPENDIX REFERENCE

Gemma Team. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*, 2025. URL https://arxiv.org/abs/2503.19786.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024. URL https://arxiv.org/abs/2310.03744.

Alec Radford et al. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, pp. 8748–8763. PMLR, 2021. URL https://proceedings.mlr.press/v139/radford21a.html.

Peng Wang et al. Qwen2-VL: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024. URL https://arxiv.org/abs/2409.12191.