GROVE: GOVERNED RULE OPERATIONALIZATION THROUGH VERIFIABLE AGENTIC EVOLUTION FOR MULTIMODAL MODERATION

Anonymous authors
Paper under double-blind review

ABSTRACT

How can a multimodal moderator adapt to an authorized policy revision without retraining its vision–language model or silently changing unrelated rules? We introduce GROVE (Governed Rule Operationalization through Verifiable Agentic Evolution), a two-stage framework that compiles policy clauses, exceptions, and scopes into a versioned schema of atomic questions. Its release protocol specifies source-bound agent edits and independent checks of policy fidelity and protected behavior. At inference, five-choice answer-token distributions supply rule-linked evidence to a lightweight decision adapter. We evaluate this evidence path on five public image–text benchmarks with group-isolated predictions and fixed-readout controls. Adding atomic evidence to a fixed XGBoost head improves macro-F1 by **3.93** points on MAMI, ROC-AUC by **1.56** points on Hateful Memes, and reject F1 by **8.69** points on HARM-C relative to the same head using only the base score. On HARM-C, false positives and false negatives decrease together. On Hateful Memes, the policy-evidence expert reaches 89.58% ROC-AUC, and a fixed-size disagreement review queue captures 252 representation-expert errors versus 207 with confidence ranking. These results show how clause-linked evidence can strengthen a lightweight moderation readout while retaining an explicit interface for policy inspection and revision.

1 INTRODUCTION

Moderation is governed by written rules, yet inference systems act on images and text. A policy owner can revise a prohibition, add an exception for quotation or counterspeech, or clarify that an image changes the meaning of otherwise harmless text. The problem is to operationalize each authorized revision without retraining the base moderator, while preserving source traceability and guarding previously correct decisions. Policy-conditioned guards demonstrate the value of explicit rules (SingGuard Team, 2026; Song et al., 2026; Piao et al., 2026); policy reasoning benchmarks show that rule interactions remain demanding (Zhang et al., 2026).

A practical system therefore needs a trace from each decision to its source rule, stable semantics for atomic questions, and an independent check before a policy revision can be released. These requirements suggest an editable layer around the model: a Policy Book of human-approved rules, exceptions, and scopes retains normative authority; a typed schema specifies what to inspect in each case; and a compact decision model uses the resulting evidence. Each component can be versioned and improved on a different timescale.

We call this approach *governed rule operationalization* and instantiate it as GROVE. An agent receives error clusters and proposes bounded edits to a source-grounded schema. A release verifier checks source fidelity, structural consistency, protected-policy behavior, and operational cost. The inference stage retrieves relevant atomic questions, scores them with a fixed vision–language model, and combines the five-choice distributions with the existing moderator. The proposed selective gate can retain the base decision, apply a policy-specific correction, or send a case for review. This base path can be supplied by the direct moderator or a policy-free representation expert, while the rule-conditioned branch retains explicit clause provenance. This design represents edits as targeted,

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.

auditable, and reversible schema operations, with release authority separate from the agent that proposes them.

GROVE connects two research questions: how to represent an authorized policy change as a checkable schema edit, and how to make its clause-level evidence useful to a moderation decision. We formulate the former as a constrained release problem and evaluate the latter on five public image–text tasks. Fixed-readout controls measure the predictive value of atomic evidence; strong representation controls and disagreement-based review examine its role alongside an existing moderation expert. This separates policy operationalization from retraining the underlying multimodal model.

Our contributions are:

- We propose a source-bound policy operationalization framework that separates normative authority, typed agent edits, and verifier-gated release, making schema revisions inspectable and reversible.
- We develop an atomic evidence interface that retains all five answer-token probabilities, question-specific semantics, routing state, and provenance for lightweight corrections and selective review.
- We demonstrate that atomic evidence improves a fixed-configuration XGBoost head by 3.93 macro-F1 points on MAMI, 1.56 ROC-AUC points on Hateful Memes, and 8.69 reject-F1 points on HARM-C. Matched strong-expert and review-queue analyses further characterize its task-specific utility.

2 Related Work

2.1 Policy-adaptive multimodal guardrails

LLaVAGuard, Llama Guard 3 Vision, and ShieldGemma 2 study image–text safeguarding (Helff et al., 2024; Chi et al., 2024; Zeng et al., 2025). SingGuard, PolicyShiftGuard, and policy-adaptive image guardrails place configurable rules closer to the model interface (SingGuard Team, 2026; Song et al., 2026; Piao et al., 2026). SafeWatch and GSPR further develop explicit policy reasoning (Chen et al., 2025; Li et al., 2025). GROVE contributes a release-oriented layer around such moderators: source clauses, executable schema versions, atomic evidence, independent checks, and reversible activation. The resulting research question concerns verified policy change, not only static classification accuracy.

2.2 Structured policy reasoning and atomic evidence

SafePyramid studies decomposed rule reasoning and dependencies (Zhang et al., 2026); R^2-Guard combines category probabilities in an editable reasoning graph (Kang & Li, 2025). RePolicy treats policy invocation as an action (Jiang et al., 2026). These approaches motivate GROVE's typed rule graph and case-specific routing. Its evidence contract also makes the five answer choices, their polarity, and missing-question states explicit, enabling downstream calibration and audit. Concept bottleneck models and post-hoc concept bottlenecks likewise expose interpretable intermediate evidence and permit concept-level interventions (Koh et al., 2020; Yuksekgonul et al., 2023). GROVE places this intermediate evidence within a policy-authority contract: edits bind to source clauses, answer semantics are versioned, and activation is separated from proposal through a release check.

2.3 Agentic repair, verification, and selective prediction

Self-Refine, CRITIC, OPRO, EvoPrompt, TextGrad, and GEPA use feedback to revise model behavior (Madaan et al., 2023; Gou et al., 2024; Yang et al., 2024; Guo et al., 2024; Yuksekgonul et al., 2024; Agrawal et al., 2026). Specification systems add executable checks and source-grounded refinement (Palumbo et al., 2026; Lu et al., 2026; Diallo et al., 2026). GROVE joins bounded agent edits with independent release tests and a protected prior behavior set. Inference-time decisions draw on probability calibration and selective prediction (Guo et al., 2017; Liu et al., 2025; Geifman & El-Yaniv, 2019; Mozannar & Sontag, 2020).

3 PROBLEM FORMULATION

3.1 SETTING AND AUTHORITY BOUNDARY

At policy-update step t, let $\mathcal{P}_t$ be a human-authorized Policy Book, $\mathcal{M}$ a base multimodal moderator whose parameters remain unchanged during the update, and $\mathcal{S}_{t-1}$ the previously released operational schema. We receive an adaptation set D_t^{tr}, a selection set D_t^{sel}, a sealed release-verification set V_t, a protected regression set R_t, and an adaptation budget B_t. The system produces a schema $\mathcal{S}_t$, an evidence adapter $\mathcal{A}_t$, and a selective gate $\mathcal{G}_t$.

We distinguish a *normative update* $\mathcal{P}_{t-1} \to \mathcal{P}_t$ from an *operational repair* $\mathcal{S}_t^{(k)} \to \mathcal{S}_t^{(k+1)}$. Only a policy owner may authorize the former. An agent may propose the latter, but it may not introduce a prohibition, exception, target, or scope without a source in $\mathcal{P}_t$. Release authority belongs to an independent verifier and, for high-impact edits, a human approver.

3.2 ATOMIC EVIDENCE AND FINAL ACTION

For case x_i, a router selects schema nodes J_i. The atomic scorer returns log probabilities ℓ_{ijc} for choices $c \in \{\mathrm{A}, \mathrm{B}, \mathrm{C}, \mathrm{D}, \mathrm{E}\}$. For a temperature τ_j fitted within the training domain,

$$q_{ij}(c; \tau_j) = \frac{\exp(\ell_{ijc}/\tau_j)}{\sum_{c'} \exp(\ell_{ijc'}/\tau_j)}. \tag{1}$$

Each node carries its own answer semantics and a mapping $w_j : \{\mathrm{A}, \ldots, \mathrm{E}\} \to [0, 1]$. For example, A can mean strong support in a support question, whereas E can mean high severity in a severity question. The full vector q_{ij} and the mapping w_j are retained; raw answer labels need not share a global A–E polarity. Before a shared readout, an explicit semantic version maps distributions to a common feature orientation. Routing, scoring failure, inapplicability, and unobservable evidence have separate states.

Let b_i be a policy-free base-path reject probability, supplied either by the direct moderator or by a representation expert, h_i the rule-conditioned adapter output, r_i reliability features, and u_i policy and input-quality metadata. The gate selects

$$a_i = \mathcal{G}_t(b_i, h_i, r_i, u_i) \in \{\mathrm{base}, \mathrm{correct}, \mathrm{defer}\}. \tag{2}$$

This action interface allows a policy update to use the existing moderator while exposing the reason and cost of each intervention.

3.3 SEQUENTIAL OBJECTIVE AND VALID RELEASE

We seek a released state that maximizes new-policy utility while penalizing operational burden:

$$\max_{\mathcal{S}_t, \mathcal{A}_t, \mathcal{G}_t} U_{\mathrm{new}} - \lambda_c \mathcal{C}_t - \lambda_s \Omega(\mathcal{S}_t) - \lambda_d \,\mathrm{DeferRate}, \tag{3}$$

subject to

$$\mathrm{Fidelity}(\mathcal{S}_t, \mathcal{P}_t) \geq \kappa, \tag{4}$$

$$\mathrm{Regret}_{\mathrm{old}}(\mathcal{S}_t, \mathcal{A}_t, \mathcal{G}_t) \leq \epsilon, \tag{5}$$

$$\mathrm{LCB}(\Delta U_t) \geq -\delta, \qquad \mathcal{C}_t \leq B_t. \tag{6}$$

Here U_{new} is the utility fixed before a policy-update episode; $\mathrm{Fidelity}$ assesses coverage of authorized clauses and exceptions; $\mathrm{Regret}_{\mathrm{old}}$ measures loss on protected prior-policy cases relative to the parent release; and $\mathcal{C}_t$ accounts for labels, model calls, and review time. The utility change ΔU_t is measured against a comparator fixed before that episode, and LCB is its one-sided lower confidence bound on V_t. Candidate screening on D_t^{sel} is not assigned this inferential interpretation: its adaptive reuse can bias a nominal interval (Dwork et al., 2015). A concrete release protocol may impose a stricter improvement margin than $-\delta$. A *valid release* satisfies the statistical constraints and hard invariants on traceability, scope, conflicts, evidence semantics, and versioning. If no candidate passes, *no-op* means that no new $\mathcal{P}_t$-compliant release is certified; it does not certify the parent schema for a changed policy. The parent can serve the changed scope only if it independently passes the new-policy checks; otherwise that scope must use an authorized fallback or human review while the previous release remains limited to its still-valid scope.

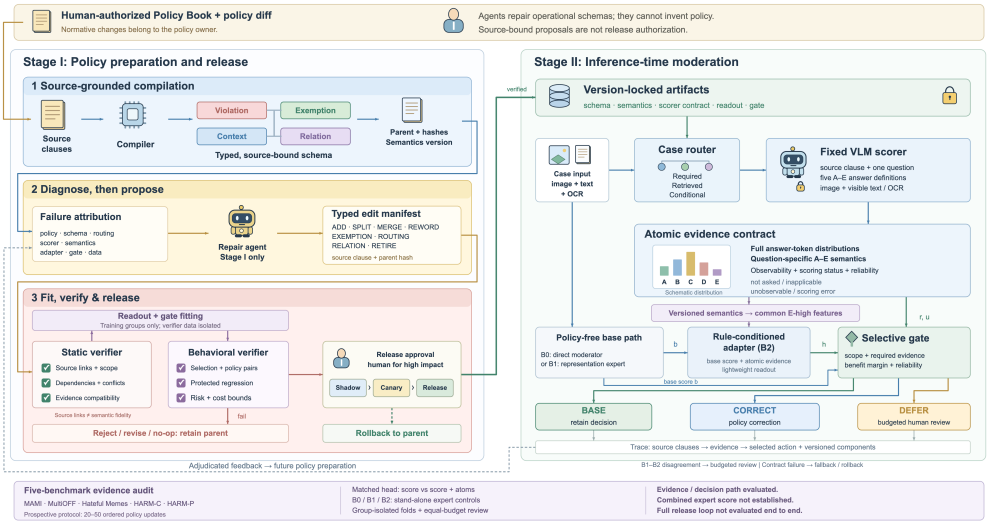

Figure 1: GROVE's two stages: policy preparation and release, followed by case-level moderation. The scorer may batch routed questions in one response.

Running example. Suppose a policy rejects derogatory claims about a protected group but exempts news reporting and counterspeech. A meme may be benign in image and text separately yet harmful through their juxtaposition. A faithful schema therefore needs nodes for the protected target, derogatory predicate, visual context, cross-modal relation, and applicable exemption. Adding a generic offensiveness node would be operationally convenient but normatively unsupported; removing the news exception to fix false negatives would violate source fidelity.

4 GROVE

4.1 Framework Overview

Figure 1 presents GROVE as a two-stage moderation workflow. Stage I prepares policy for use: it compiles the authorized Policy Book into a schema, attributes errors, proposes bounded edits, fits the lightweight decision components, verifies candidates, and releases a versioned artifact. Stage II performs case-level inference with the released components: it retrieves applicable atomic questions, elicits evidence, and makes a selective moderation decision. The evolution agent operates only in Stage I and cannot directly publish a schema.

4.2 Governed Operational Schema

To prepare an initial schema, each authorized clause is decomposed into its scope, triggering condition, exceptions, severity, and required modalities. The compiler proposes answerable questions and dependencies while preserving links to the source clauses; a cross-modal condition remains an explicit relation rather than two unrelated unimodal questions. Clause coverage and conflicts are checked before a candidate can be released.

Each schema version records its parent version, Policy Book hash, evidence-semantics version, scope, approval state, and rollback target. We summarize an atomic node as $v_j = (C_j, q_j, \sigma_j, \rho_j)$: source clauses C_j, question q_j, A–E answer semantics σ_j, and applicability and routing rules ρ_j. Node metadata captures rule kind, modality, exceptions, dependencies, and observable evidence. Required nodes always run; retrieved and conditional nodes run when their criteria hold. The full versioned contract appears in Appendix A.

Algorithm 1 Verifier-gated policy-schema evolution

Require: Authorized $\mathcal{P}_t$, parent $\mathcal{S}_{t-1}$, D_t^{tr}, D_t^{sel}, sealed V_t, R_t, budget B_t

Compile source-grounded initial schema $\mathcal{S}_t^{(0)}$
for $k = 0, 1, \ldots$ while budget remains **do**
  Execute $\mathcal{S}_t^{(k)}$ on D_t^{tr} and attribute failures
  Agent proposes a bounded set of typed edits E_k
  Reject edits that fail static verification
  Screen survivors on D_t^{sel} within the edit budget
  if a candidate passes the development screen **then**
    Set $\mathcal{S}_t^{(k+1)}$ to the best screened candidate
  else
    Retain $\mathcal{S}_t^{(k)}$; record rejection evidence
  end if
  if no admissible improvement or stopping criterion holds **then**
    break
  end if
end for
Freeze one candidate; verify once on V_t, R_t or record no new release
If verified, human approval $\rightarrow$ shadow $\rightarrow$ canary $\rightarrow$ release or rollback
return Current-policy-compliant release or no-new-release status

4.3 Failure attribution and typed evolution

Wrong decisions do not necessarily indicate missing rules. GROVE attributes an error to the authorized policy or its schema, atomic-question routing and scoring, evidence interpretation and aggregation, final-action selection, or input and label quality. This attribution determines the remedy: an omitted or misstated authorized condition motivates a schema edit, whereas routing, scoring, and decision errors are addressed in their respective components.

The evolution agent may add, split, merge, reword, or retire an atom; add an authorized exemption; or change routing or a cross-modal relation. Each typed proposal includes a parent hash, target nodes, source clauses, motivating failure clusters, expected effect, and complexity change. The edit manifest makes the candidate and its lineage inspectable; its operation names appear in Appendix A.

4.4 Independent verification

Static verifier. Deterministic checks validate schema syntax, resolvable source links, scope compatibility, and evidence-contract compatibility; they also detect cyclic dependencies, declared conflicts, and duplicate nodes. Typed metadata flags an edit that broadens a prohibition, removes an exception, or changes severity for semantic source review. High-impact changes require policy-owner approval.

Behavioral verifier. The proposer can screen candidates on D_t^{sel}. After a single candidate is frozen, an isolated verifier checks it once on sealed V_t and R_t, including same-content/different-policy pairs and predeclared high-severity slices. A failed check records no new release; another attempt requires new verification groups. The verifier reports paired uncertainty, worst-slice loss, calibration, deferral, calls, latency, and human-review cost. Sealed groups keep the final release check separate from adaptive candidate selection.

Release state machine. A candidate can be rejected, returned for revision, accepted for human approval, or left unreleased. An approved candidate enters shadow and canary stages before full release. Runtime semantic-hash mismatch, missing required evidence, abnormal harmful flips, or service failure triggers fallback to a policy-authorized base path or rollback where the parent version remains valid. An unresolved changed-policy scope is referred to human review.

4.5 Atomic evidence and calibration

The scorer receives visible text, optical character recognition (OCR) output, the image, and an ordered batch of source-linked atomic questions with their A–E meanings. A strict JSON response contains one answer per question; at each answer-token position we read the five competing log probabilities and construct equation 1. In multi-question responses these are *batch-conditional* distributions, not independently scored questions (Appendix E). The record stores question order, prompt and response-schema hashes, answer kind and polarity, routing and scoring status, entropy, margin, image hash, and model version. Full vectors retain distinctions that a single expected score can erase. When independent atomic annotations are available, temperature scaling can be fitted within the adaptation domain (Guo et al., 2017; Liu et al., 2025).

4.6 Inference-time decision adapter

The adapter aggregates a variable set of node distributions, state masks, and reliability metadata. A transparent rule combiner, logistic regression, or a shallow boosted tree can implement this interface. In our experiments, L2 logistic regression and depth-two XGBoost readouts use the same cached atomic evidence, providing compact alternatives for different task sizes.

The gate can intervene when rule scope is active, required nodes are present, and the predicted correction benefit exceeds a direction-specific margin. Ambiguous disagreements can be routed to a bounded review queue. Every decision records policy, schema, scorer, calibrator, adapter, threshold, and base-model versions with the triggered rule path. The resulting trace supports inspection of a decision and rollback to an earlier release.

5 Experimental Setup

5.1 Benchmarks and policy evidence

We evaluate MAMI (Fersini et al., 2022), MultiOFF (Suryawanshi et al., 2020), Hateful Memes (Kiela et al., 2020), and HARM-C/HARM-P (Pramanick et al., 2021). Their retrospective, group-isolated analysis cohorts contain 8,977, 440, 5,300, 2,918, and 2,900 cases, respectively. Each task uses a benchmark-specific research Policy Book as the source for its atomic catalog: 10 MAMI nodes, 10 MultiOFF nodes, 6 Hateful Memes nodes, 20 HARM-C nodes, and 17 HARM-P nodes. HARM-P asks eight required nodes and routes one of nine optional nodes per case. All five tasks use cached final-answer token log probabilities from the same Qwen3.5-35B-A3B-FP8 atomic scorer, with explicit asked/scored states. The task-specific answer semantics and feature orientation are documented in Appendix E.

The primary metrics fixed for this retrospective audit are macro-F1 for MAMI and HARM-P, reject F1 for MultiOFF and HARM-C, and ROC-AUC for Hateful Memes. We additionally report accuracy, reject precision and recall, calibration, and false-negative and false-positive rates. All paired intervals resample leakage groups 2,000 times. Each dataset is evaluated on its own metric and cohort.

5.2 Matched evidence and representation controls

The *atomic readout* compares a calibrated direct-moderation score from the same Qwen3.5 scorer (B_0), a depth-two XGBoost head trained on that score, and the same XGBoost configuration trained on the base score plus five A–E probabilities and state features for each atom. L2 logistic regression supplies a linear readout. Five outer leakage-group folds contain one prediction per case; calibration and threshold selection use training-domain nested out-of-fold signals. Learner and feature specifications are in Appendix E. Source identities, split membership, feature values, predictions, and paired statistics have integrity checks.

The *strong-control audit* adds B_1, a policy-free head over the base score and a dataset-specific joint image–text representation: Qwen3-VL-Embedding-2B for MAMI and Hateful Memes, or CLIP ViT-B/32 for MultiOFF, HARM-C, and HARM-P. Within each outer training fold, group-isolated predictions select between logistic regression and a histogram-gradient-boosting head, with training-fold principal component analysis (PCA), calibration, and threshold fitting. The policy-bearing

Table 1: Five-dataset nested out-of-fold evidence ablation. Values are the primary metric fixed for this retrospective audit, in percent. Δ compares XGBoost with atomic A–E evidence to the same XGBoost configuration with only the base score; unadjusted intervals use 2,000 paired leakage-group bootstrap replicates. Boldface marks gains with paired 95% intervals entirely above zero.

Dataset	Primary metric	B_0	Score only	Base + atoms	Δ [95% CI]
MAMI	Macro-F1	79.79	79.52	83.45	**+3.93** [3.26, 4.57]
MultiOFF	Reject F1	58.27	58.74	61.10	+2.36 [-1.07, 5.99]
Hateful Memes	ROC-AUC	87.79	87.41	88.97	**+1.56** [1.17, 1.96]
HARM-C	Reject F1	61.70	61.70	70.39	**+8.69** [6.67, 10.65]
HARM-P	Macro-F1	60.52	59.40	62.23	**+2.83** [1.10, 4.50]

B_2 is a previously developed, guarded adapter over the base score and a trained atomic-evidence expert. Its outer-training procedure selects no-op, a continuous residual correction, or a directional disagreement selector; Appendix E specifies its features, cross-fitting, calibration, and screening rule. B_0, B_1, and B_2 share case, label, leakage group, and outer-fold identities. In GROVE's decision interface, B_1 can serve as a policy-free expert alongside the rule-conditioned B_2 branch. Here B_1 remains a strong representation control: Table 2 reports their stand-alone performance; their disagreement supplies the review-selection signal evaluated below.

5.3 Policy-update evaluation design

The framework's principal evaluation unit is an authorized policy update. The proposed protocol orders 20–50 updates spanning prohibitions, exceptions, scope, severity, and cross-modal conditions. Each episode separates adaptation and selection from one-shot sealed verification, old-policy protection, and final audit. Budget-matched fixed-schema, full-policy-prompt, agent-edit, and human-edit comparators test authorized decision flips and invariant-case preservation on same-content/different-policy pairs. The design measures utility, regret, unauthorized edits, release outcomes, fidelity, and cost; Appendix D gives the prospective episode and verifier-ablation outline.

6 Results

6.1 Atomic distributions add information to a fixed learner

Table 1 compares the fixed-configuration XGBoost head with and without atomic five-choice evidence. Both variants use the same learner, training labels, group folds, calibration procedure, and threshold protocol. The atomic variant improves the primary-metric point estimate on all five tasks, with positive unadjusted paired group intervals on MAMI, Hateful Memes, HARM-C, and HARM-P. The largest matched gain is 8.69 reject-F1 points on HARM-C, followed by 3.93 macro-F1 points on MAMI and 1.56 ROC-AUC points on Hateful Memes. This matched-readout design measures the additional predictive information available in the atomic feature package, independently of changes to the learner specification.

Which decisions change. The HARM-C improvement reflects fewer errors in both directions: relative to the score-only head, atomic evidence reduces false positives from 720 to 431 (a **40.1%** reduction) and false negatives from 251 to 238. Reject precision rises from 52.06% to 64.85% while reject recall rises from 75.70% to 76.96%; ROC-AUC increases from 75.96% to 84.84%. MAMI shows the same two-sided pattern, with false positives falling from 1,142 to 857 (25.0%) and false negatives from 692 to 628. Its reject precision and recall rise by 4.96 and 1.43 points, respectively. These simultaneous error reductions and ranking gains are consistent with better discrimination, rather than a uniform shift toward more or fewer rejections.

On Hateful Memes, the atomic head increases ROC-AUC from 87.41% to 88.97% and reject precision from 71.94% to 74.84%, while reducing false positives from 546 to 474. The remaining tasks also show useful point-estimate changes: MultiOFF accuracy rises from 46.36% to 56.59% with 63 fewer false positives, and HARM-P ROC-AUC rises from 61.85% to 65.87% with 92 fewer false positives. Appendix B reports the complete operating-point metrics. Across tasks, the A–E vector

Table 2: Matched-cohort stand-alone expert results. B_1 uses joint image–text representations without policy features; B_2 is the guarded policy-evidence branch. Scores and paired differences from the calibrated base decision B_0 are in percentage points. Brackets are unadjusted 95% leakage-group bootstrap intervals (2,000 replicates); boldface marks differences whose interval excludes zero. B_2 intervals were recomputed against the same B_0 predictions used for B_1. The shared B_0 is not B_2's within-run base; Appendix E, Table 8 separates the two contrasts.

Dataset	Primary metric	B_0	B_1	B_2	$B_1 - B_0$ [95% CI]	$B_2 - B_0$ [95% CI]
MAMI	Macro-F1	79.79	86.70	84.45	**+6.91 [6.09, 7.70]**	**+4.66 [3.93, 5.40]**
MultiOFF	Reject F1	58.27	63.06	60.61	**+4.79 [1.35, 8.22]**	**+2.34 [0.81, 3.99]**
Hateful Memes	ROC-AUC	87.79	88.99	89.58	**+1.20 [0.74, 1.69]**	**+1.79 [1.37, 2.23]**
HARM-C	Reject F1	61.70	72.79	70.24	**+11.09 [8.88, 13.25]**	**+8.54 [6.50, 10.63]**
HARM-P	Macro-F1	60.52	62.44	60.72	+1.92 [-0.02, 3.81]	+0.21 [-1.47, 1.83]

supplies information beyond a single direct-moderation score while leaving the underlying vision–language scorer unchanged. Table 4 separately reports the retrospective risk screen, distinguishing predictive quality from operating-point acceptance.

6.2 Policy evidence under strong representation controls

Table 2 compares a previously developed policy-bearing adapter with a policy-free representation head on the same cases and outer folds. The guarded policy-evidence adapter B_2 improves on its own calibrated base by 4.72 macro-F1 points on MAMI, 1.79 ROC-AUC points on Hateful Memes, and 8.38 reject-F1 points on HARM-C (Table 8). These within-run contrasts distinguish the adapter's increment from differences in base calibration. MultiOFF retains the base in all five folds; its shared-B_0 contrast is therefore reported as a base-path difference. On Hateful Memes, B_2 also exceeds the policy-free representation head B_1 by 0.59 ROC-AUC points (paired 95% CI [0.09, 1.13]). Together with the matched-head ablation, this provides evidence of added value from rule-linked signals on a demanding cross-modal benchmark. The differing expert predictions motivate selective combination and the review analysis below. Table 1 measures feature-package value within a fixed learner, whereas Table 2 characterizes separately trained experts. The two comparisons address readout utility and expert-level performance, respectively.

6.3 Disagreement enriches review on Hateful Memes

We examine a fixed-budget deferral rule using disagreement between B_1 and B_2 on all five cohorts. At each dataset's disagreement-queue size, we compare the number of B_1 errors selected with a queue ranked by B_1's distance from its decision threshold (Table 5). On Hateful Memes, the disagreement rule captures 252 of B_1's errors among 470 selected cases, compared with 207 errors under an equal-size confidence queue. This improves accuracy on automatically handled cases from 84.82% to 85.76%, a paired group difference of 0.93 points [0.32, 1.50]. The disagreement queue covers 8.87% of cases while capturing 26.81% of B_1's 940 errors; the equal-size confidence queue captures 22.02%. Thus, on this task, the same number of review slots captures 45 additional errors, a **21.7%** increase in error capture. Table 5 gives the five-task comparison. This budget is the human-review count; disagreement uses predictions from both experts.

7 Analysis and Discussion

7.1 From whole-case scores to clause-level evidence

One whole-case score compresses multiple judgments about targets, actions, context, and exemptions into a single value. GROVE exposes these judgments as named question distributions and availability states that a readout can combine differently. The matched experiment demonstrates usable information in this representation: HARM-C improves both ranking and thresholded decisions, while MAMI and Hateful Memes also improve reject precision and recall together. This pattern is stronger than an improvement obtained solely by shifting a common rejection threshold.

The full A–E vector preserves distribution shape, and the state masks distinguish an unasked question from low support. Named feature groups connect readout inputs to inspectable questions. Once scored, this evidence can be cached for calibration and lightweight readout updates without further VLM calls or changes to the scorer's weights. The shallow-tree experiment therefore demonstrates a practical separation between evidence acquisition and supervised decision adaptation.

7.2 An editable interface with explicit policy authority

GROVE's design makes the editable object explicit: an agent changes a typed schema, not the governing policy or the VLM weights. Source clauses, scope, exceptions, answer meanings, and edit lineage form a contract that can be checked before activation. Separating proposal from release allows the same evidence and decision interface to accommodate human, template-based, or agent-generated repairs under a common approval boundary. Versioned artifacts also identify exactly which policy interpretation and evidence contract produced a prediction.

7.3 Complementary experts and human review

The representation head captures generic image–text features; the policy branch provides explicit clause-linked judgments. Their stand-alone scores and disagreements expose different diagnostic views of the same content. On Hateful Memes, disagreement selects an 8.87% queue containing over one-quarter of the representation expert's errors, outperforming equal-size confidence review. This supports using policy evidence for targeted review on a task where image–text interaction is central, in addition to its utility as a direct readout input.

8 Limitations

The reported evaluation covers static evidence and readout behavior, not completed agent-edit and verified-release episodes. Catalog development used these cohorts; group-isolated fitting and conditional bootstrap intervals are therefore retrospective, not a fresh blind test. Matched learners do not match question budgets: equal-cost alternatives, compressed atom features, B_1 plus atoms, and independently annotated fidelity/locality remain to be tested. B_1 has higher point estimates on four tasks, and review enrichment is task-specific. MultiOFF trades recall for specificity; three atomic-XGBoost tasks miss the retrospective acceptance screen (Table 4). Historical API measurements also require controlled end-to-end cost validation.

9 Conclusion

GROVE introduces an editable policy layer with source-bound questions, typed revisions, and a verifier-gated release design. Its evidence interface improves a fixed XGBoost readout across five image–text tasks, with positive unadjusted paired intervals on four. HARM-C gains 8.69 reject-F1 points with fewer errors in both directions; on Hateful Memes, expert disagreement concentrates errors in an 8.87% review queue. Together, the design and measured evidence gains provide a modular, traceable companion to existing multimodal moderators without retraining their VLM.

Ethics Statement

Content moderation affects expression and can distribute errors unevenly across groups. GROVE records the governing clause, applicable exception, evidence path, and release decision so these decisions can be reviewed. Normative changes remain under human policy authority. The public datasets contain hateful, misogynistic, and offensive material; annotation and access should follow their licenses and reviewer-wellbeing procedures. Authors should audit false positives on quotation, news, education, and counterspeech and false negatives on targeted harm before any deployment.

Reproducibility Statement

The appendix specifies the schema contract, cohort construction and label mapping, scorer workload, learner configurations, thresholds, paired bootstrap, and policy-update protocol. Frozen local

artifacts bind source data, image fingerprints, score journals, case identities, group folds, predictions, and metrics by hash. An artifact checker recomputes prediction identities and paired statistics; it is distinct from the proposed policy-release verifier. Anonymous release of the code, operational catalogs, and prediction records is subject to dataset licenses and must be completed for external replication.

AI USE STATEMENT

Generative AI tools assisted with literature discovery, research framing, hypothesis and experimental-protocol development, code drafting and review, qualitative diagnosis, result interpretation, and manuscript drafting and editing. The authors are responsible for the policy definitions, data use, experiments, citations, code, claims, and final manuscript. Numerical claims were checked against bound experimental artifacts.

REFERENCES

Lakshya A. Agrawal, Shangyin Tan, Dilara Soylu, Noah Ziems, Rishi Khare, Krista Opsahl-Ong, Arnav Singhvi, Herumb Shandilya, Michael J. Ryan, Meng Jiang, Christopher Potts, Koushik Sen, Alex Dimakis, Ion Stoica, Dan Klein, Matei Zaharia, and Omar Khattab. GEPA: Reflective prompt evolution can outperform reinforcement learning. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/0e9e708b6f48e14fd0ac29e167413f76-Abstract-Conference.html.

Zhaorun Chen, Francesco Pinto, Minzhou Pan, and Bo Li. SafeWatch: An efficient safety-policy following video guardrail model with transparent explanations. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/beac6bfb7eac3d651307c16ac747df01-Abstract-Conference.html.

Jianfeng Chi, Ujjwal Karn, Hongyuan Zhan, Eric Smith, Javier Rando, Yiming Zhang, Kate Plawiak, Zacharie Delpierre Coudert, Kartikeya Upasani, and Mahesh Pasupuleti. Llama Guard 3 Vision: Safeguarding human-AI image understanding conversations. *arXiv preprint arXiv:2411.10414*, 2024. URL https://arxiv.org/abs/2411.10414.

Nafi Diallo, Fernando Galves, and Ferhat Erata. Automated reasoning policy refinement in amazon bedrock. AWS Machine Learning Blog, August 2026. URL https://aws.amazon.com/blogs/machine-learning/automated-reasoning-policy-refinement-in-amazon-bedrock/. Accessed 2026-09-13.

Cynthia Dwork, Vitaly Feldman, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Aaron Roth. Generalization in adaptive data analysis and holdout reuse. *arXiv preprint arXiv:1506.02629*, 2015. URL https://arxiv.org/abs/1506.02629.

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

Yonatan Geifman and Ran El-Yaniv. SelectiveNet: A deep neural network with an integrated reject option. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 2151–2159. PMLR, 2019. URL https://proceedings.mlr.press/v97/geifman19a.html.

Zhibin Gou, Zhihong Shao, Yeyun Gong, Yelong Shen, Yujiu Yang, Nan Duan, and Weizhu Chen. CRITIC: Large language models can self-correct with tool-interactive critiquing. In *International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/fef126561bbf9d4467dbb8d27334b8fe-Abstract-Conference.html.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 1321–1330. PMLR, 2017. URL https://proceedings.mlr.press/v70/guo17a.html.

Qingyan Guo, Rui Wang, Junliang Guo, Bei Li, Kaitao Song, Xu Tan, Guoqing Liu, Jiang Bian, and Yujiu Yang. Connecting large language models with evolutionary algorithms yields powerful prompt optimizers. In *International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/9156b0f6dfa9bbd18c79cc459ef5d61c-Abstract-Conference.html.

Lukas Helff, Felix Friedrich, Manuel Brack, Kristian Kersting, and Patrick Schramowski. LlavaGuard: An open VLM-based framework for safeguarding vision datasets and models. *arXiv preprint arXiv:2406.05113*, 2024. URL https://arxiv.org/abs/2406.05113.

Houcheng Jiang, Boxuan Zhang, Qiyong Zhong, Junfeng Fang, Xiang Wang, and Xiangnan He. RePolicy: Reinforcement learning for safety-policy invocation in agent safeguards. *arXiv preprint arXiv:2608.24275*, 2026. URL https://arxiv.org/abs/2608.24275.

Mintong Kang and Bo Li. R^2-Guard: Robust reasoning enabled LLM guardrail via knowledge-enhanced logical reasoning. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/a07e87ecfa8a651d62257571669b0150-Abstract-Conference.html.

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, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/1b84c4cee2b8b3d823b30e2d604b1878-Abstract.html.

Pang Wei Koh, Thao Nguyen, Yew Siang Tang, Stephen Mussmann, Emma Pierson, Been Kim, and Percy Liang. Concept bottleneck models. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119, pp. 5338–5348. PMLR, 2020. URL https://proceedings.mlr.press/v119/koh20a.html.

Haoran Li, Yulin Chen, Jingru Zeng, Hao Peng, Huihao Jing, Wenbin Hu, Xi Yang, Ziqian Zeng, Sirui Han, and Yangqiu Song. GSPR: Aligning LLM safeguards as generalizable safety policy reasoners. *arXiv preprint arXiv:2509.24418*, 2025. URL https://arxiv.org/abs/2509.24418.

Hongfu Liu, Hengguan Huang, Xiangming Gu, Hao Wang, and Ye Wang. On calibration of LLM-based guard models for reliable content moderation. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/a99f732df9b668284b449da0214a3286-Abstract-Conference.html.

Xiaoyue Lu, Xianglin Yang, Haijun Liu, Jiahao Liu, Kuntai Cai, Yan Xiao, and Jin Song Dong. Inverting the shield: Systematically generating safety tests from policy specifications. *arXiv preprint arXiv:2605.24883*, 2026. URL https://arxiv.org/abs/2605.24883.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, Shashank Gupta, Bodhisattwa Prasad Majumder, Katherine Hermann, Sean Welleck, Amir Yazdanbakhsh, and Peter Clark. Self-Refine: Iterative refinement with self-feedback. In *Advances in Neural Information Processing Systems*, volume 36, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/91edff07232fb1b55a505a9e9f6c0ff3-Abstract-Conference.html.

Hussein Mozannar and David Sontag. Consistent estimators for learning to defer to an expert. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 7076–7087. PMLR, 2020. URL https://proceedings.mlr.press/v119/mozannar20b.html.

Nils Palumbo, Sarthak Choudhary, Jihye Choi, Guy Amir, Prasad Chalasani, and Somesh Jha. Formal policy enforcement for real-world agentic systems. *arXiv preprint arXiv:2602.16708*, 2026. URL https://arxiv.org/abs/2602.16708.

Caiyong Piao, Zhiyuan Yan, Haoming Xu, Yunzhen Zhao, Kaiqing Lin, Feiyang Xu, and Shuigeng Zhou. Towards policy-adaptive image guardrail: Benchmark and method. *arXiv preprint arXiv:2603.01228*, 2026. URL https://arxiv.org/abs/2603.01228.

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

SingGuard Team. SingGuard: A policy-adaptive multimodal LLM guardrail with dynamic reasoning. *arXiv preprint arXiv:2606.22873*, 2026. URL https://arxiv.org/abs/2606.22873.

Mingyang Song, Luxin Xu, Haoyu Sun, Minzhou Pan, Yu Cheng, and Bo Li. PolicyShiftGuard: Benchmarking and improving policy-adaptive image guardrails. *arXiv preprint arXiv:2607.05910*, 2026. URL https://arxiv.org/abs/2607.05910.

Shardul Suryawanshi, Bharathi Raja Chakravarthi, Mihael Arcan, and Paul Buitelaar. Multimodal meme dataset (MultiOFF) for identifying offensive content in image and text. In *Proceedings of the Second Workshop on Trolling, Aggression and Cyberbullying*, pp. 32–41, 2020. URL https://aclanthology.org/2020.trac-1.6/.

Chengrun Yang, Xuezhi Wang, Yifeng Lu, Hanxiao Liu, Quoc V. Le, Denny Zhou, and Xinyun Chen. Large language models as optimizers. In *International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/3339f19c5fcee3ad74502947a32be9e6-Abstract-Conference.html.

Mert Yuksekgonul, Maggie Wang, and James Zou. Post-hoc concept bottleneck models. In *International Conference on Learning Representations*, 2023. URL https://arxiv.org/abs/2205.15480.

Mert Yuksekgonul, Federico Bianchi, Joseph Boen, Sheng Liu, Zhi Huang, Carlos Guestrin, and James Zou. TextGrad: Automatic "differentiation" via text. *arXiv preprint arXiv:2406.07496*, 2024. URL https://arxiv.org/abs/2406.07496.

Wenjun Zeng, Dana Kurniawan, Ryan Mullins, Yuchi Liu, Tamoghna Saha, Dirichi Ike-Njoku, Jindong Gu, Yiwen Song, Cai Xu, Jingjing Zhou, Aparna Joshi, Shravan Dheep, Mani Malek, Hamid Palangi, Joon Baek, Rick Pereira, and Karthik Narasimhan. ShieldGemma 2: Robust and tractable image content moderation. *arXiv preprint arXiv:2504.01081*, 2025. URL https://arxiv.org/abs/2504.01081.

Jiacheng Zhang, Haoyu He, Sen Zhang, Shen Wang, Xiaolei Xu, Yuhao Sun, Meng Shen, and Feng Liu. SafePyramid: A hierarchical benchmark for in-context policy guardrailing. *arXiv preprint arXiv:2606.29887*, 2026. URL https://arxiv.org/abs/2606.29887.

A Operational Schema and Edit Contract

Versioned schema and nodes. A schema records its parent, Policy Book identity and hash, evidence semantics, scope, compiler/agent identity, approval state, and rollback target; released versions are immutable. Each node records an ID and source clause, typed subject–predicate–object relation, scope/modality, question and A–E meanings, positive/negative criteria, exceptions, severity, routing, and dependencies. The evidence contract distinguishes not asked, inapplicable, unobservable, and scoring error. Rewording must preserve or version polarity and feature semantics.

Catalog instance. The Hateful Memes research Policy Book defines a protected-characteristic attack using a human or group target, a causal link to the protected trait, and endorsement by the complete meme; quotation and rebuttal are exceptions. Its enabled atom `protected_identity_target_and_causal_link` asks whether the complete meme establishes that target and link, while `complete_meme_quotes_or_rebuts_attack` asks whether the meme quotes, reports, or rebuts rather than endorses an apparent attack. For both atoms, A means the statement is definitely satisfied, C denotes ambiguous or insufficient evidence, and E means definitely not satisfied. The catalog records the parent research Policy Book ID, source-rule IDs, criteria, and question text, illustrating how target conditions and exceptions remain individually inspectable within the same operational policy.

Typed edit manifest. Each proposal binds the parent hash, affected nodes and source clauses, operation, error cluster, rationale, expected effect, risk, complexity, verification results, and human decision. Allowed operations add, split, merge, reword, or retire atoms; add exemptions; or change routing and relations. Rejected edits are retained alongside accepted ones.

B Full Matched-Head Metrics

Table 3: Nested out-of-fold metrics for matched depth-two XGBoost heads using only the base score or the base score plus all atomic A–E features. Each case contributes one held-out prediction; reject is the positive class. Metrics are percentages.

Dataset	Features	ACC	Macro-F1	Reject F1	Reject P	Reject R	ROC-AUC	FPR
MAMI	Score	79.57	79.52	80.53	76.86	84.57	87.95	25.42
MAMI	Score + atoms	83.46	83.45	83.86	81.82	86.00	91.50	19.08
MultiOFF	Score	46.36	41.06	58.74	43.30	91.30	62.29	85.94
MultiOFF	Score + atoms	56.59	56.00	61.10	48.86	81.52	65.30	61.33
Hateful Memes	Score	80.42	78.80	72.95	71.94	74.00	87.41	16.02
Hateful Memes	Score + atoms	81.96	80.34	74.68	74.84	74.52	88.97	13.91
HARM-C	Score	66.72	66.14	61.70	52.06	75.70	75.96	38.20
HARM-C	Score + atoms	77.07	75.84	70.39	64.85	76.96	84.84	22.86
HARM-P	Score	59.41	59.40	60.12	59.77	60.46	61.85	41.66
HARM-P	Score + atoms	62.24	62.23	61.57	63.46	59.78	65.87	35.24

Within each dataset pair, the learner configuration and evaluation folds are identical; only the atomic feature vector is added. Table 1 uses the dataset-specific primary metric, while this table also displays the precision–recall profile and ranking quality. Answer probabilities are rebuilt from frozen raw logprob journals at temperature 1.0; satisfaction-descending vectors are reversed to the common E-high orientation, with a separate asked/retrieved/scored state for each node. The input has 81, 81, 49, 161, or 137 dimensions for MAMI, MultiOFF, Hateful Memes, HARM-C, or HARM-P, respectively, including the base score. The same XGBoost hyperparameters are held fixed across datasets. For fixed atom columns, A–E reversal is a bijective permutation that standardizes the interface while preserving its information content.

Table 4: Retrospective quality/risk screen. Values are one-sided 95% upper bounds (percentage points) for candidate minus comparator FNR/FPR from 2,000 paired leakage-group bootstrap resamples. The retrospective run-manifest screen also requires primary-metric gain ≥ 0.5 points and a positive unadjusted two-sided 95% lower bound; each rate bound must be $\leq +1$ point. The Table 1 XGBoost contrast uses its score-only head as reference. B_2 uses its own within-run calibrated base, not Table 2's separately calibrated B_0. These are retrospective acceptance diagnostics, not prospective policy-release decisions.

	Atomic XGBoost vs score only			B_2 vs within-run base		
Dataset	FNR UB	FPR UB	Pass	FNR UB	FPR UB	Pass
MAMI	-0.68	-5.48	yes	-1.16	-6.43	yes
MultiOFF	$+14.59$	-19.77	no	0.00	0.00	no
Hateful Memes	$+0.48$	-1.41	yes	-0.71	-2.13	yes
HARM-C	$+1.20$	-13.54	no	$+0.86$	-12.97	yes
HARM-P	$+2.74$	-4.33	no	$+2.79$	-1.09	no

Table 5: Equal-size review queues on all five cohorts. Captured errors count B_1 mistakes among cases selected for review. Automatic accuracy is computed on the remaining cases. The disagreement queue uses $B_1 \neq B_2$; the confidence queue selects the same number of cases closest to B_1's decision threshold. VLM inference costs are not matched.

Dataset	Queue size	Captured errors		Automatic ACC (%)	
		Disagree	Confidence	Disagree	Confidence
MAMI	1,109	454	472	90.59	90.82
MultiOFF	111	47	47	56.23	56.23
Hateful Memes	470	**252**	207	**85.76**	84.82
HARM-C	487	192	202	83.79	84.20
HARM-P	891	420	440	66.75	67.75

C Failure Taxonomy and Oracle Protocol

For each error, the diagnostic record stores the active policy and schema versions, routed and missing nodes, node distributions, input-quality flags, base and adapter scores, final action, gold decision, and reviewer adjudication. Oracle routing supplies the policy-relevant node set without altering node scores. Oracle atomic scoring supplies independent node truth without changing routing. Oracle aggregation receives correct routed evidence. Oracle selection chooses between already-produced base and plugin decisions. Each oracle is computed only when its required independent annotation is available; its output diagnoses a failure layer and is excluded from deployable manifests.

D Policy-Update Evaluation Outline

Each authorized policy diff fixes the parent schema, edit and feedback budgets, grouped adaptation/selection data, sealed verification, old-policy protection, and untouched audit cases. Independent reviewers label same-content cases under both policies, including expected flips in both directions and invariant cases. Budget-matched controls include no update, full-policy prompting, unverified or human edits, and verifier ablations. The planned metrics are new-policy utility, old-policy regret, unauthorized edits, release/no-op/rollback outcomes, fidelity, and cost. Only one frozen candidate reaches sealed verification; audit labels are revealed after its decision. Mechanism controls compare equal-cost alternative questions, full versus compressed A–E evidence, and B_1 versus B_1 plus atoms. Independent clause and case–node labels are needed to assess semantic fidelity and scorer accuracy separately.

E RETROSPECTIVE PROTOCOL DETAILS

Catalog and split provenance. The fixed catalogs contain 10 MAMI, 10 MultiOFF, 6 Hateful Memes, 20 HARM-C, and 17 HARM-P active atoms; HARM-P routes nine per case. They were developed using these cohorts before the reported readout audit, so outer-fold isolation applies to model fitting. Leakage groups are connected by an identical image SHA-256 or by a caption SHA-256 after NFKC normalization, case folding, and punctuation collapse; transitive matches stay in one outer fold.

Table 6: Construction of the five retrospective readout cohorts. "Source" counts labeled source-table rows considered by the adapter; "usable" removes unavailable images or missing labels before scoring. "Train" is the usable source-training split. The nested evaluation uses only the stricter `selection_train` cohort; groups and positives refer to that cohort.

Dataset	Source	Usable	Train	Evaluated	Groups	Positive
MAMI	11,000	11,000	9,000	8,977	8,898	4,485
MultiOFF	743	740	444	440	438	184
Hateful Memes	9,540	7,575	6,744	5,300	4,458	1,892
HARM-C	3,544	3,544	3,013	2,918	2,750	1,033
HARM-P	3,470	3,469	2,937	2,900	2,842	1,467

Positive labels are MAMI Task-A, MultiOFF `Offensive`, the provided Hateful Memes label, and HarMeme `somewhat/very harmful`. Exclusions are three unavailable MultiOFF images, 1,965 unavailable Hateful Memes images, and one missing HARM-P label. Source-split leakage groups are purged before evaluation; Hateful Memes additionally reserves one of five group folds for selection. These are retrospective training-source cohorts, not official blind test scores.

For Hateful Memes, 1,965/9,540 source rows (20.60%) lack a local image: 1,756 train, 102 dev-seen, and 107 dev-unseen rows. Their positive-label rates are 35.59%, 48.04%, and 36.45%, respectively, versus 35.50%, 49.75%, and 37.18% among retained rows in the same splits. This observable label check does not establish equivalence in image content; the evaluated 5,300 cases come only from rows with available images.

The direct-head audit uses five group-isolated outer folds and three inner folds for calibration and thresholds. Learners are L2 logistic regression ($C = 0.1$) or XGBoost (depth 2, 200 trees, learning rate 0.03, minimum child weight 20, $\ell_2 = 10$). Training-only Brier selection chooses identity, one-temperature, or nonnegative-slope Platt calibration. Held-out labels are used only for evaluation; paired intervals use 2,000 group-bootstrap replicates. Atomic evidence uses temperature 1.0 and catalog-specific A–E semantics, oriented to E-high for the shared feature interface.

Observed scoring workload. Table 7 summarizes the successful request journals for exactly the evaluated cases. MAMI, MultiOFF, and HARM-C batch their atoms in one response; Hateful Memes and HARM-P score separately. Each row also includes one base response. Tokens and API latency sum successful responses per case, then report the median; retries, parallelism, and human-review cost are not included. HARM-C's historical latency is service-dependent, not a controlled estimate of schema cost.

Table 7: Historical VLM scoring workload for evaluated cases. Atomic and total successful responses are separate columns; each total includes one base response. Actual API attempts may be higher. Tokens and API latency are median per-case sums including that base response, so it must not be added again.

Dataset	Active atoms	Atomic responses	All responses	Total tokens	API latency (s)
MAMI	10	1	2	2,243	6.6
MultiOFF	10	1	2	2,444	6.2
Hateful Memes	6	6	7	5,241	35.0
HARM-C	20	1	2	2,146	401.1
HARM-P	9 per case	9	10	4,865	22.6

Batch-conditional token evidence. MAMI, MultiOFF, and HARM-C use one autoregressively generated, flat JSON object per case; they do not use independent server-side sequences for each atom. Hateful Memes and HARM-P use one atomic question per response. For the jth answer in a multi-atom response, the measured vector is conditional on the entire ordered batch and the previously generated JSON prefix:

$$q_{ij}^{\mathrm{batch}}(c) \propto \exp(\ell_{ijc}), \quad \ell_{ijc} = \log P_\theta\Big(Y_{ij} = c \mid x_i, Q_{i,1:m}, \hat{Y}_{i,<j}, \text{JSON prefix}\Big). \tag{7}$$

This describes the observation process, not independence across atoms. Requests use temperature zero, a strict JSON schema with A–E string values, returned token log probabilities, and the top 20 alternatives. The parser matches each JSON value span to its emitted token position, accepts exact choice-token variants with surrounding whitespace or quotes, and takes all five candidate log probabilities *at that position and prefix*. It normalizes over A–E only when all five candidates are present. A missing candidate, unparseable response, or alignment failure invalidates the response for retry; none is silently filled with zero probability. Every successful atomic response in the frozen journals has five observed candidates per atom. The ordered policy IDs, prompt hash, response-schema hash, catalog hash, and selected token index identify the scoring context.

For a desensitized MAMI journal example, the request contains ten ordered policy IDs, shared A–E definitions, the visible text and image (omitted here), and `top_logprobs=20`. Its strict schema requires those ten IDs as keys, disallows other keys, and restricts each value to A–E. The first two answers, for overall meaning and gendered target, are B at token 17 and A at token 33. At token 17 the A–E log probabilities are $(-1.352, -1.102, -2.602, -1.852, -1.727)$; their five-way normalized feature is $(.259, .332, .074, .157, .178)$. The remaining answers are extracted at their own positions in the same response. These are model features, not gold labels.

Replay boundary. Rebuilding the reported tables requires the versioned research Policy Books and complete atom catalogs (questions, A–E meanings, routing, and source links), base and atomic log-probability journals, case/label/group lists, matched representations, and B_2 fold records. Training-only fits and 2,000 paired group-bootstrap draws reconstruct Tables 1– 8. Hashes identify files but cannot replace accessible artifacts; anonymous release remains subject to dataset licenses.

The representation control selects logistic regression or histogram gradient boosting within each outer training fold, using the base score and Qwen3-VL or CLIP joint embedding. Logistic candidates use standardized features, $C \in \{0.01, 0.1\}$, and optional class balancing. Boosting uses training-only randomized PCA (at most 64 components), 160 iterations, 15 leaves, 20 minimum samples per leaf, learning rate 0.05, and $\ell_2 = 1$. Three inner group folds select the learner; PCA, calibration, and thresholds never use outer-held-out labels.

Within each outer fold, B_2 combines calibrated whole-case Qwen scores with a policy expert (HGB-L2, except XGBoost on Hateful Memes). Its previously chosen evidence temperatures are $2, 2, 1, 1, 2.5$ in dataset order. Five-fold group cross-fitting supplies expert predictions to an L2 logistic residual ($C = 0.03$) using calibrated logits, their difference and magnitude, expert decisions, named atoms, and missingness indicators. A disagreement selector is trained only with at least 100 disagreement rows and 20 cases per target class. Three-fold training-only meta-predictions calibrate the stack and choose its policy; outer-held-out labels are never used to fit it.

The policy menu contains no-op, residual thresholds 0.10–0.90 in 0.01 steps, and one-/two-direction disagreement selectors with cost ratios $1, 2, 5$. Eligibility requires at least 0.005 training-domain

primary gain with neither FP nor FN count increasing; ties favor fewer changes. Nineteen of 25 outer folds selected a residual, six no-op, and none a selector. Thus B_2 is a binary expert, not the three-way gate; the review queue and bootstrap risk screen are separate retrospective analyses.

Table 8 separates the policy branch from the two calibrated base paths. For each task metric M, $M(B_2) - M(B_{0,\text{shared}})$ equals $[M(B_2) - M(B_{0,\text{within}})] + [M(B_{0,\text{within}}) - M(B_{0,\text{shared}})]$. These are point-estimate identities, not causal decompositions of the underlying metric. MultiOFF's five no-op folds make its within-run increment exactly zero. The nonzero shared-base contrast on that task therefore cannot be credited to policy evidence.

Table 8: Decomposition of B_2's stand-alone comparison. Values are each task's primary metric in percent; increments are percentage points computed before rounding. "Within" is the calibrated base path fitted in B_2's own run. Residual/no-op counts cover five outer folds per task.

Dataset	Shared B_0	Within B_0	B_2	B_2-within	Within$-$shared	Residual/no-op
MAMI	79.79	79.73	84.45	+4.72	-0.06	5/0
MultiOFF	58.27	60.61	60.61	0.00	+2.34	0/5
Hateful Memes	87.79	87.79	89.58	+1.79	0.00	5/0
HARM-C	61.70	61.85	70.24	+8.38	+0.15	5/0
HARM-P	60.52	59.75	60.72	+0.97	-0.76	4/1

The $B_0/B_1/B_2$ comparison binds identical case, label, group, and fold identities. Artifact checks recompute metrics, split disjointness, and hashes; they are not the policy-release verifier. Table 2 uses paired group-bootstrap intervals (2,000 resamples for $B_2 - B_0$); ROC-AUC uses probabilities and threshold metrics use persisted decisions. Table 5 uses these predictions, breaking queue ties by case ID.