RUBRIC CAN HALLUCINATE TOO: DIAGNOSING AND REPAIRING MULTIMODAL RUBRIC HALLUCINATION

Anonymous authors
Paper under double-blind review

ABSTRACT

As multimodal large language models (MLLMs) are deployed in increasingly open-ended settings, scalar correctness and holistic preference scores provide limited diagnostic feedback. Evaluation and alignment are therefore shifting toward rubric-based assessment, which decomposes response quality into explicit, scoreable criteria. However, these criteria are not always reliable: they may demand information absent from the image, assert unverifiable motives, or require deliverables beyond what the task asks for. We term this phenomenon **Multimodal Rubric Hallucination (MRH)** and formalize rubric validity along two axes: *Evidence Support* and *Task Fit*. Unlike conventional multimodal hallucination in model responses, MRH makes the evaluation standard itself the object of diagnosis. We propose a new task, **MRH Diagnosis and Repair**: identify one of six MRH types and rewrite the hallucinated rubric into an evidence-supported, task-aligned replacement. We introduce **MRH-Bench**, comprising 8,472 verified multimodal instances organized into three progressively challenging tracks: *Rubric hallucination type identification*, *Hallucinated rubric repair*, and *LLM-Human evaluation agreement assessment*. Across eight representative proprietary and open-weight MLLMs, the strongest prompted baseline reaches only 61.0% Macro-F1 and 15.80% repair success. We further propose **MRH-Guard**, a two-stage framework combining structured weighted supervised fine-tuning with **Quadrant-Aware Policy Optimization (QAPO)** to strengthen the diagnostic decisions that guide repair. MRH-Guard reaches 75.4% Macro-F1 (+14.4 points) and 19.20% repair success (+3.40 points).

1 INTRODUCTION

The more open-ended and high-stakes the task, the less a single score can say about whether a multimodal large language model (MLLM) answer is acceptable (Chen et al., 2024; Kumar et al., 2026; Hu et al., 2026). Recent multimodal evaluation therefore moves beyond holistic judgments toward rubric-level assessment, where task fulfillment, factual grounding, and perceptual quality can be examined separately (Lee et al., 2024b; Kim et al., 2024; Wei et al., 2026). Rubrics operationalize this shift by providing explicit criteria for model-based judges (Liu et al., 2023; Zheng et al., 2023; Lee et al., 2024b) and structured signals for multimodal reward modeling and policy optimization (Kan et al., 2026; Tian et al., 2026). Both rest on one assumption: *the rubric itself is reliable*.

This assumption is fragile. Model-assisted rubric construction is increasingly common (Kan et al., 2026; Tian et al., 2026), yet seemingly precise criteria may exceed the available evidence or task scope. In Figure 1(a), a clothing-description task is paired with a rubric requiring the response to mention a silver zipper that is not visible in the image. This requirement can reward unsupported details rather than faithful task completion.

We term this phenomenon **Multimodal Rubric Hallucination (MRH)**. A rubric exhibits MRH when a scoring requirement lacks *Evidence Support*, meaning it cannot be verified from the image and task-authorized references, or *Task Fit*, meaning it exceeds the prompt's task scope. Whereas prior work primarily diagnoses hallucinations in model responses (Li et al., 2023; Cai et al., 2025; Jin et al., 2026), MRH makes the evaluation standard itself the object of diagnosis. An invalid requirement can therefore systematically bias downstream judgments and reward signals.

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.

Evaluating MRH poses three challenges (Figure 1). First, diagnosis requires evidence and task context: a plausible requirement may lack visual support or fall outside the prompt's scope. Second, rubric errors can distort downstream evaluation: a judge following an unsupported requirement may reward hallucinated details and penalize grounded responses, producing judgments that conflict with human judgments. Third, repair must preserve valid evaluative intent and be validated downstream: removing an unsupported requirement is insufficient if useful criteria are lost, and a plausible rewrite does not establish that the same judge will agree more closely with humans.



Figure 1: Three challenges in evaluating and repairing multimodal rubric hallucinations.

We address these challenges with a new task, **MRH Diagnosis and Repair**, which requires a model to identify the type of MRH and rewrite the hallucinated rubric while preserving its grounded evaluative intent. Our six-type taxonomy captures both what multimodal rubrics assess and how their requirements can exceed the task scope or available evidence. We construct **MRH-Bench** with 8,472 multimodal instances across three progressive tracks: 3,472 for Rubric hallucination type identification, 3,000 for Hallucinated rubric repair, and 2,000 for LLM-Human evaluation agreement assessment. Our three-stage construction pipeline first curates eligible source items, then generates controlled MRH instances, repairs, and human judgments, and finally subjects the resulting examples to independent model and expert verification. Only instances approved by both verifier models and all three expert reviewers are retained.

On MRH-Bench, we evaluate eight representative proprietary and open-weight MLLMs as prompted baselines. The strongest prompted baseline reaches only 61.0% Macro-F1 in MRH type identification and 15.80% joint repair success. We therefore propose **MRH-Guard**, a two-stage framework for MRH diagnosis and repair. It first learns type identification and repair through weighted supervised fine-tuning, and then strengthens the diagnostic decisions that guide repair using **Quadrant-Aware Policy Optimization (QAPO)**. MRH-Guard improves over the strongest prompted baseline by 14.4 Macro-F1 points and 3.40 percentage points in repair success, while increasing downstream Spearman's ρ by 0.0135. Our contributions are threefold:

- We introduce, for the first time, **Multimodal Rubric Hallucination** as a distinct failure mode of multimodal evaluation. We formalize criterion validity through Evidence Support and Task Fit, and formulate **MRH Diagnosis and Repair**, which requires models to identify the failure mechanism and produce an evidence-supported, task-aligned replacement.
- We derive a six-type taxonomy of MRH mechanisms and construct **MRH-Bench**, comprising 8,472 independently verified multimodal instances across MRH type identification, rubric repair, and downstream score alignment.
- We propose **MRH-Guard**, which combines structured weighted supervised fine-tuning with QAPO to improve MRH diagnosis, repair, and downstream judge–human agreement, outperforming the strongest baselines by 14.4% on Macro-F1, 3.40% in repair success.

2 Related Work

Multimodal Hallucination Diagnosis. Multimodal hallucination research has largely treated hallucination as a property of model responses. Li et al. (2023) probe object existence with polling

Table 1: Six types of Multimodal Rubric Hallucination

Type	Definition
T1	**Task-Scope Overreach.** Adds a goal, deliverable, or hard constraint outside the task scope.
T2	**Unsupported Visual Claims.** Requires unsupported claims about directly observable content, such as objects, attributes, spatial relations, or text.
T3	**Unwarranted Inference.** Turns observations into unsupported conclusions about intention, emotion, causality, future events, or social meaning.
T4	**Reference Overconstraint.** Makes nonessential reference wording, format, or details mandatory scoring conditions.
T5	**Unsupported External Facts.** Requires exact factual details needing external evidence not provided or authorized by the task.
T6	**Modality/Provenance Overreach.** Requires unavailable sensory information or unsupported claims about origin, authenticity, or capture metadata.

questions, Guan et al. (2024) construct image–text control groups that entangle language hallucination with visual illusion, Sun et al. (2023) score hallucination rate and response informativeness with MMHal-Bench, and Wang et al. (2023) cover existence, attribute, and relation hallucinations without an LLM judge. More recent work moves from binary detection toward fine-grained diagnosis: Cai et al. (2025) localize hallucinated spans at the token level, while Jin et al. (2026) jointly detect response-level hallucinations, classify their types, and generate corrective explanations. Despite this increasing granularity, the response remains the object of diagnosis and the evaluation criterion is treated as fixed. MRH shifts the object of diagnosis to the criterion itself: it asks whether the criterion is evidence-supported and task-aligned, identifies how it fails, and repairs it before use.

Rubric-guided Multimodal Evaluation. A second line of work makes evaluation criteria explicit through checklists, rubrics, and learned judges (Zheng et al., 2023; Lee et al., 2024b; Kim et al., 2024). Multimodal judges vary across scoring and ranking settings and exhibit task-dependent uncertainty and perceptual bias (Chen et al., 2024; Kumar et al., 2026; Hu et al., 2026; Park et al., 2026). Recent methods improve judge generalization with multitask reinforcement learning (Wu et al., 2026), calibrate perceptual evaluation with atomic visual criteria (Wei et al., 2026), and use rubrics for generative reward modeling or visual-faithfulness optimization (Zhang et al., 2026; Kan et al., 2026; Tian et al., 2026). Related optimization methods target verifiable outcomes, multimodal perception, or response hallucination (Shao et al., 2024; Wang et al., 2026; Li et al., 2026b). These studies improve judges or responses under the assumption that the given rubrics are valid. However, a hallucinated rubric is a more fundamental failure: it can systematically misguide evaluation, reward modeling, and model optimization. This motivates MRH-Bench, which treats rubric validity as the prediction target. To the best of our knowledge, MRH-Bench is the first benchmark for rubric hallucination. A structured comparison with existing benchmarks is provided in Appendix B.

3 THE MRH-BENCH

3.1 TAXONOMY

We derive the MRH taxonomy from two complementary mappings in multimodal evaluation, i.e., **what** a rubric may legitimately evaluate and **how** a rubric requirement can become unsupported.

From multimodal capabilities to rubric design directions. Existing VQA and MLLM benchmarks assess instruction following, visual and document understanding, reasoning, reference use, and external knowledge (Liu et al., 2024; Yue et al., 2024; Masry et al., 2022). These capabilities provide the basis for rubric design: each criterion operationalizes a capability required by the task as an explicit, scoreable requirement. The prompt defines the task scope, the provided modalities define the observable evidence, and authorized references or knowledge define what additional information may legitimately be required.

From multimodal response hallucinations to rubric hallucinations. Multimodal hallucination studies organize response errors around nonexistent objects, unsupported attributes and relations,

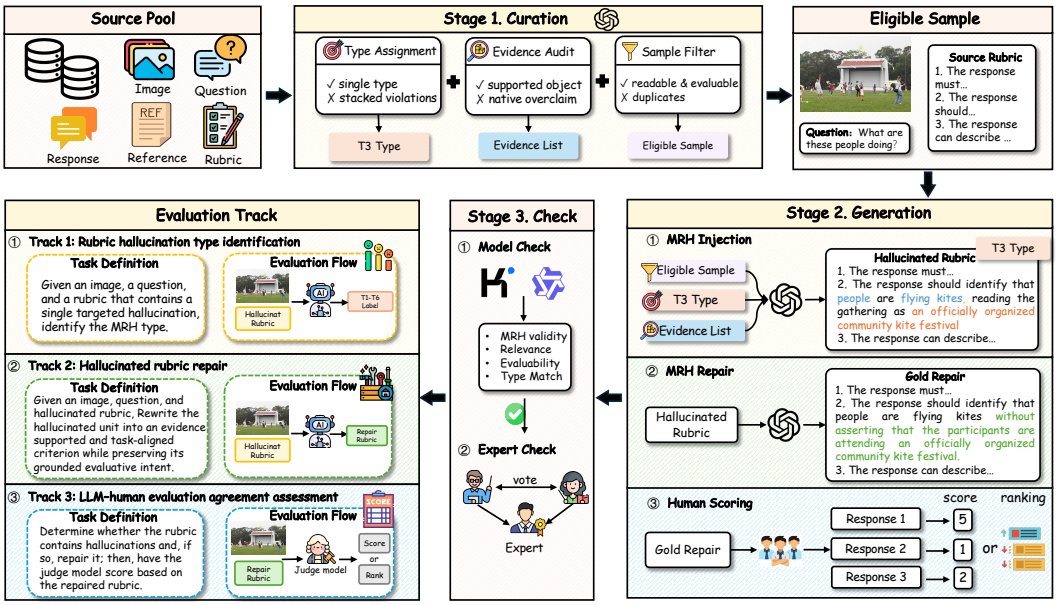


Figure 2: Construction pipeline and evaluation tracks of MRH-BENCH.

reasoning overclaims, and unverifiable knowledge (Li et al., 2023; Wang et al., 2023; Guan et al., 2024; Cai et al., 2025). However, this response-centric view assumes that the evaluation rubrics themselves are valid, overlooking cases in which a rubric requires unsupported content. We therefore extend hallucination analysis from model responses to evaluation rubrics: a rubric hallucinates when it requires a response to make or reproduce an unsupported claim.

Synthesis. Combining the two mappings yields six categories designed to be mutually exclusive: Task-Scope Overreach (T1), Unsupported Visual Claims (T2), Unwarranted Inference (T3), Reference Overconstraint (T4), Unsupported External Facts (T5), and Modality/Provenance Overreach (T6). Table 1 summarizes their definitions. Detailed category boundaries and annotation rules are provided in Appendix E.

3.2 DATASET CONSTRUCTION

As shown in Figure 2, we build MRH-BENCH in three stages: curating eligible source items, generating a controlled MRH instance and its blind repair, and independently verifying both outputs.

Source pool. We collect data from nine primary sources: MLLM-as-a-Judge (Chen et al., 2024), Perception-Collection (Lee et al., 2024a), MM-RLHF (Zhang et al., 2025), GenExam (Wang et al., 2025), WaferMap-VQA (Xu & Lian, 2026), MIA-Bench (Qian et al., 2024), MMTutorBench (Yang et al., 2025), FACTS-Multimodal (Cheng et al., 2025), and GRADE (Liu et al., 2026), together with smaller auxiliary sources summarized as *Others* in Figure 3. Collectively, they cover natural images, charts, diagrams, documents, scientific figures, and image editing tasks.

Stage 1: Curation. We use GPT-5.6 Sol (OpenAI, 2026b) to perform three checks. *Sample filtering* removes corrupted, incomplete, underspecified, and duplicate items. *Evidence audit* identifies the observable visual evidence relevant to each source rubric. *Type confirmation* retains items that admit a controlled injection of one designated MRH type without changing the prompt or other valid requirements. Each retained item is paired with an evidence list and a target type.

Stage 2: Generation. GPT-5.6 Sol is invoked twice with separate contexts. **MRH Injection.** In the first call, the model receives the eligible sample, evidence list, target type, and source rubric. It rewrites one criterion to introduce the assigned MRH type while preserving the original topic and valid requirements, producing the hallucinated rubric used across the three tracks. **MRH Repair.** In

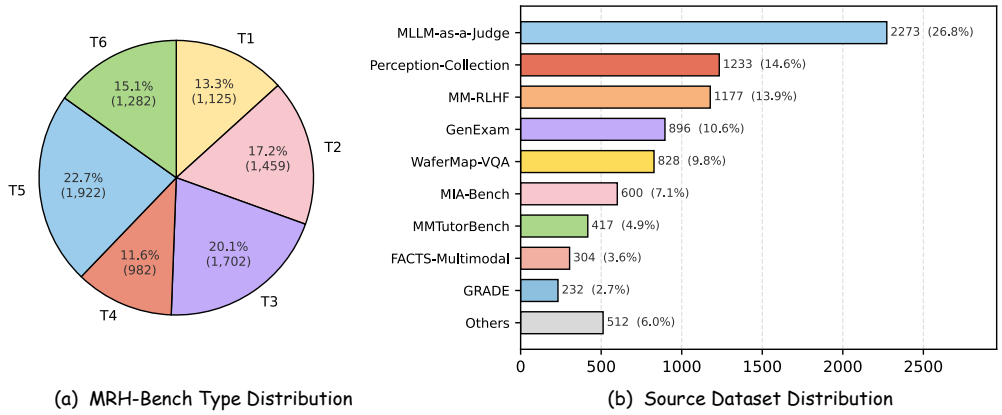


Figure 3: Distribution of the 8,472 instances in MRH-BENCH.

the second call, the same model receives only the image, prompt, and hallucinated rubric; the source rubric and target type are hidden. It positively rewrites the affected scoring unit into an evidence-supported and task-aligned gold repair while preserving the evaluative intent. **Human Scoring or Ranking.** Candidate responses are withheld from both model calls, three annotators score or rank them under the gold repair to produce the references; their reliability is reported in Table 8.

Stage 3: Check. Kimi-K3 (Kimi Team, 2026) and Qwen3.8-Max (Qwen Team, 2026c) independently review each hallucinated-rubric–repair pair without access to the construction rationale. They check MRH validity, task relevance, type match, and evaluability for the rubric, and MRH resolution, intent preservation, and absence of new unsupported content for the repair. Pairs approved by both models undergo independent review by three experts using the same seven fields (Tables 9 and 10). Only consensus-approved pairs are retained after adjudication; rejected pairs are regenerated or removed. Appendix D reports the review protocol and pre-adjudication Fleiss' κ. Figure 3 summarizes the final distribution.

3.3 EVALUATION TRACKS

We build a progressive evaluation chain with three tracks to isolate three capabilities, i.e., identifying the type of a hallucinated rubric, repairing the affected criterion, and preserving human-aligned judgments when the repair is used for downstream scoring.

Track 1: Rubric hallucination type identification. Track 1 contains 3,472 verified hallucinated rubrics. Each instance provides the image, prompt, hallucinated rubric, and any reference or exemplar explicitly authorized by the source task. The model must predict the single MRH type in $\{\mathrm{T1}, \ldots, \mathrm{T6}\}$. Appendix M.1 gives two representative instances. Because every item contains exactly one verified target mechanism, Track 1 evaluates fine-grained type identification rather than binary screening. We report Macro-F1 and Weighted-F1.

Track 2: Hallucinated rubric repair. Track 2 contains 3,000 verified instances paired with gold repairs. Given the task evidence and hallucinated rubric, the model generates a repaired replacement. A valid repair positively rewrites the complete affected scoring unit, preserves its grounded evaluative intent, and introduces no unsupported requirement. We report residual MRH (R-MRH), grounded-content preservation (G-Pres), newly introduced unsupported content (New-Unsup), semantic fidelity (Sem-Fid), and joint repair success (RS) rather than lexical overlap; detailed definitions and computation are provided in Appendix A, and two representative cases in Appendix M.2.

Track 3: LLM–Human evaluation agreement assessment. Track 3 contains 2,000 cases with held-out candidate responses and a hidden human reference per case: 731 for absolute scoring and 1,269 for ranking. At evaluation, the tested model generates a repair, and a fixed downstream judge evaluates the same responses under it. The task measures agreement between these induced deci-

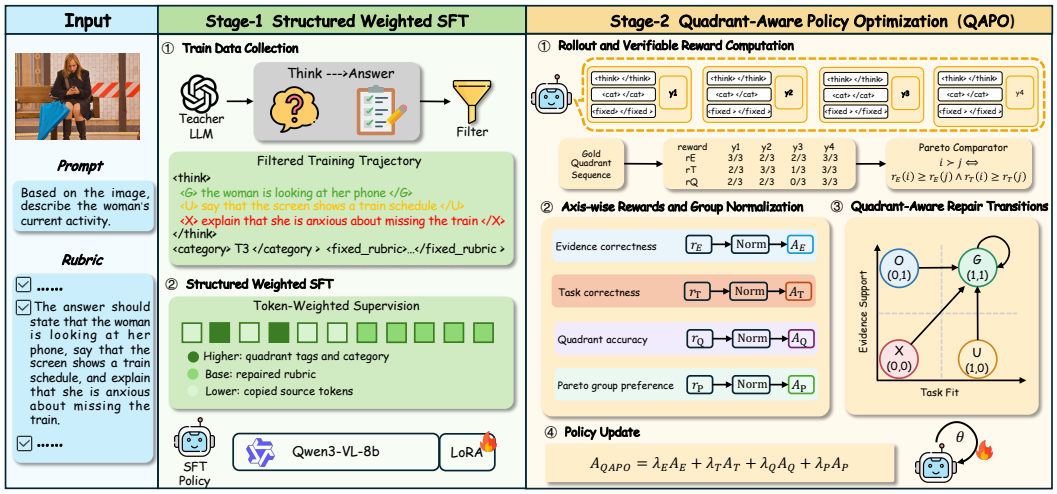


Figure 4: Overview of the two-stage MRH-GUARD training framework.

sions and the human references, not textual similarity to the gold repair. We report Spearman's ρ, QWK, and MAE for absolute scoring, and PairAcc and Kendall's τ_b for ranking; detailed definitions and computation are provided in Appendix A, and four representative cases in Appendix M.3.

4 METHOD: MRH-GUARD

We formulate **MRH Diagnosis and Repair** as identifying the hallucination mechanism and producing an evidence-supported, task-aligned rubric. **MRH-Guard** first learns structured diagnosis and repair through weighted supervised fine-tuning (SFT), then uses Quadrant-Aware Policy Optimization (QAPO) to refine Evidence Support and Task Fit decisions (Figure 4).

4.1 TASK FORMULATION

Given $x = (I, P, c, R)$, comprising an image, prompt, hallucinated rubric, and optional reference, the model predicts an MRH type $y \in \{\mathrm{T1}, \ldots, \mathrm{T6}\}$ and repaired criterion c^*:

$$z = \langle \texttt{think} \rangle\, g\, \langle \texttt{/think} \rangle\, \langle \texttt{category} \rangle\, y\, \langle \texttt{/category} \rangle \\ \langle \texttt{fixed_rubric} \rangle\, c^*\, \langle \texttt{/fixed_rubric} \rangle. \tag{1}$$

Here, g is a unit-state-tagged copy of c, with no free-form reasoning.

We segment c into M complete, independently scoreable units $\{s_m\}_{m=1}^{M}$, each assigned binary Task Fit t_m and Evidence Support e_m. G/O/U denote *Grounded*, *Out-of-scope*, and *Unsupported*; X fails both axes. These states guide repair, whereas T1–T6 identifies the MRH mechanism; the two label spaces are not one-to-one. With coordinates ordered as (t_m, e_m), c^* preserves G and positively rewrites O/U/X rather than deleting them:

$$\begin{aligned} \mathrm{G}(1,1) &\xrightarrow{\text{preserve}} \mathrm{G}(1,1), & \mathrm{O}(0,1) &\xrightarrow{\text{repair Task Fit}} \mathrm{G}(1,1), \\ \mathrm{U}(1,0) &\xrightarrow{\text{repair Evidence}} \mathrm{G}(1,1), & \mathrm{X}(0,0) &\xrightarrow{\text{repair both}} \mathrm{G}(1,1). \end{aligned} \tag{2}$$

4.2 STAGE 1: STRUCTURED WEIGHTED SFT

Each training target is the structured tuple (g_i, y_i, c_i^*). Because most tokens in g_i copy the source rubric, uniform token loss would underweight the diagnostic decisions. We instead optimize

$$\mathcal{L}_{\mathrm{WSFT}}(\theta) = -\frac{1}{N} \sum_{i=1}^{N} \frac{\sum_t w_{i,t} \log \pi_\theta(z_{i,t} \mid x_i, z_{i,<t})}{\sum_t w_{i,t}}, \\ w_{\mathrm{diag}} = w_{\mathrm{cat}} = \alpha, \quad w_{\mathrm{repair}} = \beta, \quad w_{\mathrm{copy}} = 1, \qquad \alpha > \beta > 1. \tag{3}$$

The four weights apply to G/O/U/X tags, T1–T6 labels, repair tokens, and copied or structural tokens, respectively. Because the loss is normalized by the total token weight, we fix $w_{\text{copy}} = 1$ without loss of generality. Exact token masks, weight values, and sensitivity analyses are reported in Appendix K. Stage 1 establishes explicit diagnosis and repair before policy optimization.

4.3 Stage 2: Quadrant-Aware Policy Optimization

Although Stage 1 supervises all four quadrant states, token-level likelihood does not explicitly optimize axis-wise or joint diagnostic accuracy at the sequence level. Initialized from the SFT policy, QAPO samples a group of K outputs $\mathcal{G}(x) = \{z^{(i)}\}_{i=1}^{K}$. A deterministic parser aligns their tags to the M gold units and extracts $(\hat{t}_m^{(i)}, \hat{e}_m^{(i)})$. Let $q_m = (t_m, e_m)$ and $\hat{q}_m^{(i)} = (\hat{t}_m^{(i)}, \hat{e}_m^{(i)})$. With $\mathrm{Acc}_M(\hat{u}, u) = M^{-1} \sum_m \mathbb{1}[\hat{u}_m = u_m]$, we define

$$r_E(i) = \mathrm{Acc}_M(\hat{e}^{(i)}, e), \quad r_T(i) = \mathrm{Acc}_M(\hat{t}^{(i)}, t), \quad r_Q(i) = \mathrm{Acc}_M(\hat{q}^{(i)}, q). \tag{4}$$

The first two rewards provide axis-wise partial credit, whereas r_Q rewards their exact conjunction.

To favor joint improvement, i Pareto-dominates j ($i \succ j$) when it is no worse on r_E and r_T and strictly better on at least one. Its group-relative dominance balance is

$$r_P(i) = \frac{1}{K-1} \sum_{j \neq i} \left(\mathbb{1}[i \succ j] - \mathbb{1}[j \succ i] \right). \tag{5}$$

Unlike r_Q, which evaluates each unit independently, r_P favors non-compensatory joint improvements relative to other rollouts in the same group. Unlike scalar-reward GRPO (Shao et al., 2024), QAPO normalizes each channel within the rollout group before combining advantages:

$$\begin{aligned} A_a(i) &= \frac{r_a(i) - \mu_a^{\mathcal{G}}}{\sigma_a^{\mathcal{G}} + \epsilon}, \quad a \in \{E, T, Q, P\}, \\ A_{\text{QAPO}}(i) &= \lambda_E A_E(i) + \lambda_T A_T(i) + \lambda_Q A_Q(i) + \lambda_P A_P(i), \end{aligned} \tag{6}$$

where $\lambda_a \geq 0$ and $\sum_a \lambda_a = 1$. We optimize the standard clipped group-relative objective and regularize against the frozen Stage 1 policy. QAPO directly scores the diagnostic prefix; Tracks 2 and 3 test whether its sequence-level credit transfers to repair quality and downstream scoring. Channel weights, numerical stabilization, and optimization details are given in Appendix K; state-transition fidelity and transition-conditioned advantage assignment are analyzed in Appendix J.

5 Experiments

5.1 Experimental Setup

Baseline models. We compare MRH-Guard with eight representative multimodal models under a unified prompting and evaluation protocol. Seven are API-based models: GPT-5.5 (OpenAI, 2026a), Claude Sonnet 4.6 (Anthropic, 2026b), Gemini 3.7 Flash (Doshi, 2026), Qwen3.8-Flash-Next (Qwen Team, 2026a), GLM-5.3-Flash (GLM-5-Team, 2026), Kimi-K2.6 (Moonshot AI, n.d.), and MiniMax-M3 (MiniMax, 2026). We also include Qwen3-VL-8B-Instruct (Bai et al., 2025), the backbone of MRH-Guard, and two trained controls. To assess cross-backbone generalization, we evaluate MRH-Guard on Qwen3.5-9B (Qwen Team, 2026b) across all three tracks (Appendix G). All methods receive identical task evidence and track-specific instructions. Track 1 uses deterministic parsing, while Tracks 2–3 use Claude Fable 5 (Anthropic, 2026a) as judge. Additional evaluations with GPT-6 Astra (OpenAI, 2026c), including human calibration and matched-subset comparisons, are reported in Appendix H.

Evaluation metrics. For Track 1, Macro-F1 is the primary metric and Weighted-F1 measures performance under the empirical type distribution. For Track 2, repair success (RS) is the primary metric: a repair succeeds only if it removes the original MRH, preserves grounded content, introduces no unsupported content, and remains semantically faithful to the intended scoring target. We also report each condition separately as R-MRH, G-Pres, New-Unsup, and Sem-Fid; all Track 2 values are percentages. For Track 3, Spearman's ρ is primary for absolute scoring, with quadratic weighted

Table 2: Track 2 results. RS denotes the conjunction of no residual MRH, grounded-content preservation, no newly introduced unsupported content, and semantic fidelity.

Method	R-MRH (%) ↓	G-Pres (%) ↑	New-Unsup (%) ↓	Sem-Fid (%) ↑	RS (%) ↑
Prompted baselines					
GPT-5.5	9.93	**33.10**	**19.70**	19.23	15.80
Claude Sonnet 4.6	12.10	30.67	39.60	16.27	12.80
Qwen3.8-Flash-Next	18.17	26.90	23.30	16.80	13.20
GLM 5.3 Flash	16.17	27.37	28.90	18.63	12.60
Gemini 3.7 Flash	15.80	28.93	23.60	13.60	11.50
Kimi-K2.6	16.50	27.00	29.07	17.00	12.80
MiniMax-M3	17.90	28.80	37.23	16.07	10.50
Qwen3-VL-8B	37.10	20.53	47.07	10.67	5.30
Trained Qwen3-VL-8B variants					
SFT	12.40	30.20	28.50	18.30	15.60
SFT+GRPO	8.90	29.70	26.40	20.10	17.60
MRH-GUARD	**6.63**	29.17	25.60	**21.67**	**19.20**

kappa (QWK) and mean absolute error (MAE) as complementary measures. Pairwise accuracy (PairAcc) is primary for ranking, and Kendall's τ_b measures full-order consistency. Mathematical definitions and failure handling are given in Appendix A; training configurations, implementation details, and weight selection are detailed in Appendix K.

5.2 MAIN RESULTS

Track 1: Rubric hallucination type identification. Table 3 reports six-way classification performance. MRH-GUARD reaches 0.754 Macro-F1 and 0.765 Weighted-F1, outperforming the strongest prompted baseline, GPT-5.5, by 14.4 and 13.7 points, respectively. Its 1.1-point gap between Weighted-F1 and Macro-F1 is also smaller than GPT-5.5's 1.8-point gap, suggesting less sensitivity to the empirical type distribution. The Qwen3-VL-8B backbone obtains only 0.391 Macro-F1, showing that the gain cannot be attributed to the backbone.

Table 3: Track 1 results. Higher is better; best results are in bold.

Method	Macro-F1 ↑	Weighted-F1 ↑
Prompted baselines		
GPT-5.5	0.610	0.628
Claude Sonnet 4.6	0.587	0.591
Qwen3.8-Flash-Next	0.556	0.561
GLM 5.3 Flash	0.522	0.538
Gemini 3.7 Flash	0.498	0.509
Kimi-K2.6	0.473	0.491
MiniMax-M3	0.447	0.461
Qwen3-VL-8B	0.391	0.413
Trained Qwen3-VL-8B variants		
SFT	0.705	0.729
SFT+GRPO	0.741	0.747
MRH-GUARD	**0.754**	**0.765**

Track 2: Hallucinated rubric repair. Table 2 evaluates whether a generated criterion is a valid drop-in replacement rather than merely a fluent rewrite. MRH-GUARD achieves the highest RS at 19.20%, outperforming GPT-5.5, the strongest prompted baseline on this primary metric, by 3.40 percentage points, or 21.5% relatively. It also yields the lowest residual hallucination rate (6.63%) and the highest semantic fidelity (21.67%), improving over GPT-5.5 by 3.30 and 2.43 percentage points, respectively. GPT-5.5 remains stronger on grounded-content preservation and avoidance of newly introduced unsupported content: it obtains 33.10% G-Pres and 19.70% New-Unsup, compared with 29.17% and 25.60% for MRH-GUARD, advantages of 3.93 and 5.90 percentage points, respectively. Thus, MRH-GUARD's primary gain comes from resolving the targeted violation while preserving the intended scoring semantics on the same instance, rather than uniformly dominating every marginal repair property.

Track 3: LLM-Human evaluation agreement assessment. Table 4 measures how closely a fixed judge using repaired rubrics agrees with human judgments. For absolute scoring, MRH-GUARD obtains the highest Spearman correlation (0.6255) and QWK (0.6180), together with the lowest MAE (0.7240). Relative to the strongest prompted baseline in each column, these correspond to gains

Table 4: Track 3 results, averaged over three runs. Green values show absolute improvements of MRH-GUARD over the strongest prompted baseline in each column.

Method	Absolute Scoring						Ranking		
	Spearman $\rho \uparrow$		QWK $\uparrow$		MAE $\downarrow$		PairAcc $\uparrow$	Kendall $\tau_b \uparrow$	
Prompted baselines									
GPT-5.5	0.6120		0.5940		0.7649		**0.7720**	0.4806	
Claude Sonnet 4.6	0.6080		0.6000		0.7680		0.7637	0.4800	
Qwen3.8-Flash-Next	0.5904		0.5821		0.7853		0.7631	0.4687	
GLM 5.3 Flash	0.5990		0.5796		0.8060		0.7650	0.4510	
Gemini 3.7 Flash	0.6007		0.5890		0.7795		0.7608	0.4646	
Kimi-K2.6	0.5820		0.5740		0.7950		0.7460	0.4470	
MiniMax-M3	0.5680		0.5590		0.8120		0.7340	0.4290	
Qwen3-VL-8B	0.4072		0.4388		0.8953		0.5575	0.3587	
Trained Qwen3-VL-8B variants									
SFT	0.5862		0.5780		0.7930		0.7410	0.4520	
SFT+GRPO	0.6083		0.6020		0.7570		0.7584	0.4740	
MRH-GUARD	**0.6255**	+0.0135	**0.6180**	+0.0180	**0.7240**	+0.0409	0.7715	**0.4937**	+0.0131

of 0.0135 in Spearman and 0.0180 in QWK, together with a 0.0409 reduction in MAE. For ranking, MRH-GUARD achieves the highest Kendall τ_b at 0.4937 and essentially matches GPT-5.5 in PairAcc (0.7715 versus 0.7720). These results show that the Track 2 improvement transfers beyond the wording of the repaired rubric to the decisions made when that rubric is used for evaluation.

Cross-track trend. MRH-GUARD consistently outperforms its untuned Qwen3-VL-8B backbone across all three tracks, while the relative performance of prompted baselines varies, particularly in Track 2. Their differing trade-offs among hallucination removal, content preservation, and unsupported additions suggest that type identification, valid repair, and downstream alignment are related but non-equivalent capabilities, supporting their separate evaluation.

5.3 ABLATION STUDIES

Table 5: Ablation results across the three tracks.

Variant	Macro-F1 $\uparrow$	RS (%) $\uparrow$	$\rho \uparrow$	PairAcc $\uparrow$
MRH-GUARD	**0.754**	**19.20**	**0.6255**	**0.7715**
SFT Only	0.705	15.60	0.5862	0.7410
SFT+GRPO	0.741	17.60	0.6083	0.7584
w/o r_E	0.729	16.93	0.6021	0.7528
w/o r_T	0.724	16.47	0.5986	0.7497
w/o r_Q	0.713	16.07	0.5924	0.7441
w/o r_P	0.748	17.93	0.6138	0.7622

Table 5 ablates both training stages with the same backbone and training split; all Stage 2 variants use matched rollout and update budgets, while SFT Only omits Stage 2. Relative to SFT Only, the full model gains 0.049 Macro-F1, 3.60 RS points, 0.0393 Spearman's ρ, and 0.0305 PairAcc, demonstrating the value of Stage 2. Replacing QAPO with SFT+GRPO reduces all four metrics, supporting channel-wise normalization. Each reward removal also hurts performance: removing the joint-quadrant reward r_Q causes the largest overall decline, whereas removing the Pareto reward r_P has the smallest effect. These results suggest that the four rewards are complementary and that improved diagnosis transfers to repair and downstream alignment; axis-level results are provided in Appendix I.

6 CONCLUSION

We introduce **Multimodal Rubric Hallucination** (MRH), formalize rubric validity through Evidence Support and Task Fit, and formulate **MRH Diagnosis and Repair**. We construct **MRH-Bench**, an independently verified benchmark of 8,472 instances spanning type identification, rubric repair, and downstream score alignment. We further propose **MRH-Guard**, which combines structured weighted SFT with QAPO's axis-wise, joint-quadrant, and Pareto rewards. MRH-Guard improves over the strongest prompted baseline by 14.4 Macro-F1 points and 3.40 percentage points in repair success, while increasing downstream Spearman correlation by 0.0135. These results establish MRH as a distinct failure surface in multimodal evaluation.

AI USE STATEMENT

We used LLM-based tools to polish the grammar, clarity, and concision of the manuscript. We also used GPT-5.6 Sol to generate synthetic hallucinated rubrics and candidate repairs from existing multimodal source samples for MRH-BENCH, following the protocol in Section 3.2. The generated pairs underwent independent model verification and expert review (Appendix D). The authors retain full responsibility for the final text, claims, and data.

ETHICS STATEMENT

This work is guided by the ICLR Code of Ethics. We use existing multimodal datasets, with human annotators providing benchmark verification and evaluation references. Reuse and redistribution of source material remain subject to the original licenses and applicable privacy requirements. We recognize potential biases in source data and model-based judgments, and use independent model verification and expert review to reduce annotation errors. Our framework is intended to improve evaluation reliability and support, rather than replace, human judgment. We are committed to transparency and integrity throughout the research process.

REPRODUCIBILITY STATEMENT

Section 3 describes benchmark construction, with source composition, category boundaries, and verification protocols detailed in Appendices C, E, and D. Section 4 specifies the model's output format and training objectives; Appendix K reports data splits, optimization settings, hardware, and training costs. Metric definitions and evaluation prompts are provided in Appendices A and L, while Appendix H documents judge calibration and cross-judge comparisons. API-based evaluations may vary with model updates and stochastic inference.

REFERENCES

Anthropic. Claude Fable 5 and Claude Mythos 5. Anthropic, June 2026a. URL https://www.anthropic.com/news/claude-fable-5-mythos-5.

Anthropic. Introducing Claude Sonnet 4.6. Anthropic, February 2026b. URL https://www.anthropic.com/news/claude-sonnet-4-6. Accessed: 2026-09-26.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-VL technical report. *arXiv preprint arXiv:2511.21631*, 2025.

Yishuo Cai, Renjie Gu, Jiaxu Li, Xuancheng Huang, Junzhe Chen, Xiaotao Gu, and Minlie Huang. MHALO: Evaluating MLLMs as fine-grained hallucination detectors. In *Findings of the Association for Computational Linguistics: ACL 2025*, 2025.

Dongping Chen, Ruoxi Chen, Shilin Zhang, Yinuo Liu, Yaochen Wang, Huichi Zhou, Qihui Zhang, Yao Wan, Pan Zhou, and Lichao Sun. MLLM-as-a-Judge: Assessing multimodal LLM-as-a-Judge with vision-language benchmark. In *International Conference on Machine Learning*, 2024.

Qiaosheng Chen, Yang Liu, Lei Li, Kai Chen, Qipeng Guo, Gong Cheng, and Fei Yuan. InteractScience: Programmatic and visually-grounded evaluation of interactive scientific demonstration code generation. *arXiv preprint arXiv:2510.09724*, 2025.

Aileen Cheng, Alon Jacovi, Amir Globerson, Ben Golan, Charles Kwong, Chris Alberti, Connie Tao, Eyal Ben-David, Gaurav Singh Tomar, Lukas Haas, Yonatan Bitton, et al. The FACTS leaderboard: A comprehensive benchmark for large language model factuality. *arXiv preprint arXiv:2512.10791*, 2025.

Domenic V. Cicchetti and Alvan R. Feinstein. High agreement but low kappa: II. Resolving the paradoxes. *Journal of Clinical Epidemiology*, 43(6):551–558, 1990.

Tulsee Doshi. Introducing Gemini 3.7 Flash. Google Blog, August 2026. URL https://blog.google/innovation-and-ai/models-and-research/gemini-models/introducing-gemini-3-7-flash/. Accessed: 2026-09-26.

Alvan R. Feinstein and Domenic V. Cicchetti. High agreement but low kappa: I. The problems of two paradoxes. *Journal of Clinical Epidemiology*, 43(6):543–549, 1990. doi: 10.1016/0895-4356(90)90158-L.

GLM-5-Team. Glm-5: from vibe coding to agentic engineering, 2026. URL https://arxiv.org/abs/2602.15763.

Tianrui Guan, Fuxiao Liu, Xiyang Wu, Ruiqi Xian, Zongxia Li, Xiaoyu Liu, Xijun Wang, Lichang Chen, Furong Huang, Yaser Yacoob, Dinesh Manocha, and Tianyi Zhou. HallusionBench: An advanced diagnostic suite for entangled language hallucination and visual illusion in large vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

Guangzheng Hu, Ziyue Jiang, Weixu Qiao, Lixin Zhang, Jianye Kang, Yuru Wu, Rong Bao, Niantong Li, Wei Wang, Ziyi Cheng, Xinfa Zhu, Hangrui Hu, Ting He, Bing Zhao, Lin Qu, Hu Wei, and Jin Xu. OmniJudge or OmniBias? diagnosing multimodal judges through balanced, decoupled lenses. *arXiv preprint arXiv:2608.24160*, 2026.

Weilin Jin, Mingyu Wang, Wenbo Li, Hao-Yang Huang, Yifan Wu, Ying Li, Gang Huang, and Zhonghai Wu. Halluscope: Fine-grained hallucination diagnosis for multimodal large language models. *ArXiv*, abs/2607.21105, 2026. URL https://api.semanticscholar.org/CorpusID:290467100.

Zijian Kan, Wei Wang, Long Luo, Bing Zhao, Xuan Ren, Weixu Qiao, Wenbo Li, Hu Wei, and Lin Qu. RubricRM: Generative reward modeling via dynamic rubrics for image generation and editing. *arXiv preprint arXiv:2608.26956*, 2026.

Seungone Kim, Juyoung Suk, Shayne Longpre, Bill Yuchen Lin, Jamin Shin, Sean Welleck, Graham Neubig, Moontae Lee, Kyungjae Lee, and Minjoon Seo. Prometheus 2: An open source language model specialized in evaluating other language models. *arXiv preprint arXiv:2405.01535*, 2024.

Kimi Team. Kimi k3: Open frontier intelligence. 2026. URL https://api.semanticscholar.org/CorpusID:290625162.

Terry K. Koo and Mae Y. Li. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of Chiropractic Medicine*, 15(2):155–163, 2016. doi: 10.1016/j.jcm.2016.02.012.

Divake Kumar, Sina Tayebati, Devashri Naik, Ranganath Krishnan, and Amit Ranjan Trivedi. VLM judges can rank but cannot score: Task-dependent uncertainty in multimodal evaluation. *arXiv preprint arXiv:2604.25235*, 2026.

J. Richard Landis and Gary G. Koch. The measurement of observer agreement for categorical data. *Biometrics*, 33(1):159–174, 1977. doi: 10.2307/2529310.

Seongyun Lee, Seungone Kim, Sue Hyun Park, Geewook Kim, and Minjoon Seo. Prometheus-Vision: Vision-language model as a judge for fine-grained evaluation. *arXiv preprint arXiv:2401.06591*, 2024a.

Yukyung Lee, Joonghoon Kim, Jaehee Kim, Hyowon Cho, Jaewook Kang, Pilsung Kang, and Najoung Kim. CheckEval: A reliable LLM-as-a-Judge framework for evaluating text generation using checklists. *arXiv preprint arXiv:2403.18771*, 2024b.

Bo Li, Yida Yin, Wenhao Chai, Xingyu Fu, and Zhuang Liu. UEval: A benchmark for unified multimodal generation. *arXiv preprint arXiv:2601.22155*, 2026a.

Lei Li, Yuancheng Wei, Zhihui Xie, Xuqing Yang, Yifan Song, Peiyi Wang, Chenxin An, Tianyu Liu, Sujian Li, Bill Yuchen Lin, Lingpeng Kong, and Qi Liu. VL-RewardBench: A challenging benchmark for vision-language generative reward models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 2023.

Yuanshuai Li, Yuping Yan, Jirui Han, Fei Ming, Lingjuan Lv, and Yaochu Jin. IRIS: Implicit reward-guided internal sifting for mitigating multimodal hallucination. In *Proceedings of the 43rd International Conference on Machine Learning*, 2026b. URL https://arxiv.org/abs/2602.01769.

Mingxin Liu, Ziqian Fan, Zhaokai Wang, Leyao Gu, Zirun Zhu, Yiguo He, Yuchen Yang, Changyao Tian, Xiangyu Zhao, Ning Liao, Shaofeng Zhang, Qibing Ren, Zhihang Zhong, Xuanhe Zhou, Junchi Yan, and Xue Yang. GRADE: Benchmarking discipline-informed reasoning in image editing. *arXiv preprint arXiv:2603.12264*, 2026.

Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, Kai Chen, and Dahua Lin. MMBench: Is your multi-modal model an all-around player? In *European Conference on Computer Vision*, 2024.

Yuxuan Liu, Tianchi Yang, Shaohan Huang, Zihan Zhang, Haizhen Huang, Furu Wei, Weiwei Deng, Feng Sun, and Qi Zhang. Calibrating llm-based evaluator. *arXiv preprint arXiv:2309.13308*, 2023.

Ahmed Masry, Do Xuan Long, Jia Qing Tan, Shafiq Joty, and Enamul Hoque. ChartQA: A benchmark for question answering about charts with visual and logical reasoning. In *Findings of the Association for Computational Linguistics: ACL 2022*, 2022.

MiniMax. MiniMax M3: Frontier coding, 1M context, native multimodality — all in one model. MiniMax Research, June 2026. URL https://www.minimax.io/blog/minimax-m3. Accessed: 2026-09-26.

Moonshot AI. Kimi K2.6: From code to creation, from one to many. Kimi, n.d. URL https://www.kimi.ai/ai-models/kimi-k2-6. Accessed: 2026-09-26.

OpenAI. Introducing GPT-5.5. OpenAI, April 2026a. URL https://openai.com/index/introducing-gpt-5-5/. Accessed: 2026-09-26.

OpenAI. Previewing GPT-5.6 Sol: A next-generation model. OpenAI, June 2026b. URL https://openai.com/index/previewing-gpt-5-6-sol/.

OpenAI. GPT-6 Astra: A new generation of intelligence. OpenAI, 2026c. URL https://openai.com/index/gpt-6-astra/. Accessed: 2026-09-26.

Seojeong Park, Jiho Choi, Junyong Kang, Seonho Lee, Jaeyo Shin, and Hyunjung Shim. Mitigating perceptual judgment bias in multimodal LLM-as-a-judge via perceptual perturbation and reward modeling. In *Proceedings of the 43rd International Conference on Machine Learning*, 2026. URL https://icml.cc/virtual/2026/poster/61108.

Shu Pu, Yaochen Wang, Dongping Chen, Yuhang Chen, Guohao Wang, Qi Qin, Zhongyi Zhang, Zhiyuan Zhang, Zetong Zhou, Shuang Gong, Yi Gui, Yao Wan, and Philip S. Yu. Judge anything: MLLM as a judge across any modality. *arXiv preprint arXiv:2503.17489*, 2025.

Yusu Qian, Hanrong Ye, Jean-Philippe Fauconnier, Peter Grasch, Yinfei Yang, and Zhe Gan. MIA-Bench: Towards better instruction following evaluation of multimodal LLMs. *arXiv preprint arXiv:2407.01509*, 2024.

Qwen Team. Qwen3.8-Flash-Next: A new architecture, towards ultimate cost-efficiency, August 2026a. URL https://qwen.ai/blog?id=qwen3.8-flash-next.

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

Qwen Team. Qwen3.8-Max: A new bar for coding and cowork, August 2026c. URL https://qwen.ai/blog?id=qwen3.8.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. DeepSeekMath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

Zhiqing Sun, Sheng Shen, Shengcao Cao, Haotian Liu, Chunyuan Li, Yikang Shen, Chuang Gan, Liang-Yan Gui, Yu-Xiong Wang, Yiming Yang, Kurt Keutzer, and Trevor Darrell. Aligning large multimodal models with factually augmented RLHF. *arXiv preprint arXiv:2309.14525*, 2023.

Shulin Tian, Minglun Li, Yuhao Dong, Hao Ding, Jiarui Yao, Haiwen Diao, Jingkang Yang, Hongyuan Zhu, and Ziwei Liu. V-Rubrics: Visual faithfulness via rubric-based reinforcement learning. *arXiv preprint arXiv:2608.25580*, 2026.

Junyang Wang, Yuhang Wang, Guohai Xu, Jing Zhang, Yukai Gu, Haitao Jia, Jiaqi Wang, Haiyang Xu, Ming Yan, Ji Zhang, and Jitao Sang. AMBER: An LLM-free multi-dimensional benchmark for MLLMs hallucination evaluation. *arXiv preprint arXiv:2311.07397*, 2023.

Zhaokai Wang, Penghao Yin, Xiangyu Zhao, Changyao Tian, Yu Qiao, Wenhai Wang, Jifeng Dai, and Gen Luo. GenExam: A multidisciplinary text-to-image exam. *arXiv preprint arXiv:2509.14232*, 2025.

Zhenhailong Wang, Xuehang Guo, Sofia Stoica, Haiyang Xu, Hongru Wang, Hyeonjeong Ha, Xiusi Chen, Yangyi Chen, Ming Yan, Fei Huang, and Heng Ji. Perception-aware policy optimization for multimodal reasoning. In *International Conference on Learning Representations*, 2026.

Yana Wei, Hongbo Peng, Yanlin Lai, Liang Zhao, Kangheng Lin, En Yu, Keyu Lv, Han Zhou, Yin Tang, Haodong Li, Mitt Huang, Hangyu Guo, Jianjian Sun, Zheng Ge, Xiangyu Zhang, Daxin Jiang, and Vishal M. Patel. PerceptionRubrics: Calibrating multimodal evaluation to human perception. In *Proceedings of the 43rd International Conference on Machine Learning*, 2026. URL https://icml.cc/virtual/2026/poster/61890.

Junjie Wu, Xuan Kan, Zihao He, Shunwen Tan, Bo Pan, and Kaitai Zhang. Multi-task reinforcement learning for enhanced multimodal LLM-as-a-judge. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 6: Industry Track)*, pp. 1829–1847, 2026. doi: 10.18653/v1/2026.acl-industry.126.

Tianyi Xiong, Yi Ge, Min Li, Zuolong Zhang, Pranav Kulkarni, K. J. Wang, Qing He, Zeying Zhu, Chenxi Liu, Ruibo Chen, Tong Zheng, Y. Chen, Xiyao Wang, Renrui Zhang, Wenhu Chen, and Heng Huang. Multi-Crit: Benchmarking multimodal judges on pluralistic criteria-following. *arXiv preprint arXiv:2511.21662*, 2025.

Ke Xu and Zhongyuan Lian. WaferSAGE: Large language model-powered wafer defect analysis via synthetic data generation and rubric-guided reinforcement learning. *arXiv preprint arXiv:2604.27629*, 2026.

Tengchao Yang, Sichen Guo, Mengzhao Jia, Jiaming Su, Yuanyang Liu, Zhihan Zhang, and Meng Jiang. MMTutorBench: The first multimodal benchmark for AI math tutoring. *arXiv preprint arXiv:2510.23477*, 2025.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, Cong Wei, Botao Yu, Ruibin Yuan, Renliang Sun, Ming Yin, Boyuan Zheng, Zhenzhu Yang, Yibo Liu, Wenhao Huang, Huan Sun, Yu Su, and Wenhu Chen. MMMU: A massive multi-discipline multimodal understanding and reasoning benchmark for expert AGI. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

Jingyu Zhang, Kun Yang, Ming Wen, Jiawei Zhao, Yuxuan Liu, Zhuoer Xu, and Shiwen Cui. GenAlign: Towards unified alignment framework of MLLMs via generative reward model. In *Proceedings of the 43rd International Conference on Machine Learning*, 2026. URL https://icml.cc/virtual/2026/poster/62209.

Yi-Fan Zhang, Tao Yu, Haochen Tian, Chaoyou Fu, Peiyan Li, Jianshu Zeng, Wulin Xie, Yang Shi, Huanyu Zhang, Junkang Wu, Xue Wang, Yibo Hu, Bin Wen, Fan Yang, Zhang Zhang, Tingting Gao, Di Zhang, Liang Wang, Rong Jin, and Tieniu Tan. MM-RLHF: The next step forward in multimodal LLM alignment. *arXiv preprint arXiv:2502.10391*, 2025.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena. In *Advances in Neural Information Processing Systems*, 2023.

A Evaluation Metrics and Protocol

For Track 1, let $C = 6$, N_c be the support of MRH type c, and $F1_c$ its one-vs-rest F1 score. We report $\text{Macro-F1} = C^{-1}\sum_c F1_c$ and $\text{Weighted-F1} = N^{-1}\sum_c N_c F1_c$. An unparseable prediction is assigned to an invalid class and is therefore incorrect.

For each Track 2 repair i, an independent evaluator returns four binary decisions: h_i indicates residual MRH, p_i grounded-content preservation, n_i newly introduced unsupported content, and f_i semantic fidelity to the intended scoring target. Over N cases, we compute

$$\begin{aligned} \text{R-MRH} &= \frac{1}{N}\sum_i h_i, & \text{G-Pres} &= \frac{1}{N}\sum_i p_i, \\ \text{New-Unsup} &= \frac{1}{N}\sum_i n_i, & \text{Sem-Fid} &= \frac{1}{N}\sum_i f_i, \\ \text{RS} &= \frac{1}{N}\sum_i (1-h_i)p_i(1-n_i)f_i. \end{aligned} \tag{7}$$

Thus, RS credits a repair only when it resolves the hallucination, preserves the grounded evaluative content, introduces no unsupported requirement, and remains faithful to the intended scoring target. An unparseable repair receives $(h_i, p_i, n_i, f_i) = (1, 0, 1, 0)$. Evaluator or API failures are retried and no Track 2 result is reported until every case has a valid evaluator decision.

For Track 3, the fixed judge directly assigns criterion-wise and overall scores on the human 0–5 scale, and the metrics use its overall score. Spearman's ρ is computed over all held-out responses in the 731 absolute-scoring cases. QWK rounds both the judge score and the averaged human reference to the nearest integer; MAE uses both directly. For ranking cases, PairAcc micro-averages all response pairs with a strict human preference; a predicted tie receives half credit. We compute Kendall's τ_b within each case and macro-average it across the 1,269 ranking cases, assigning zero when a candidate ordering is entirely tied. Track 3 metrics are reported only when every selected case has complete fixed-judge outputs under the model-generated repair.

B Comparison with Existing Benchmarks

Table 6 distinguishes MRH-Bench from representative multimodal hallucination and evaluation benchmarks by their primary object and task interface. Response-hallucination benchmarks determine whether model outputs are grounded, whereas judge benchmarks measure how well an evaluator scores or ranks responses under supplied rubric. Neither setting directly treats the validity of those rubric as the prediction target. MRH-Bench instead requires rubric-level type identification and repair, then applies the repaired rubric to held-out responses to test whether it restores agreement with human judgments.

C Dataset Composition and Source Tasks

Table 7 summarizes the source datasets used to construct MRH-Bench. We use hallucination errors observed in their rubrics as seeds and expand the dataset through controlled perturbations anchored in the original tasks and available evidence. The benchmark spans general multimodal judging, preference alignment, instruction following, factuality, visual question answering, and domain-specific visual reasoning. For compactness, the 512-instance long tail is summarized by five representative auxiliary sources. Counts are reported separately for each track and sum to 8,472 track-specific instances.

Table 6: Comparison with representative multimodal hallucination and evaluation benchmarks. "Downstream validation" means evaluating held-out responses under a model-repaired rubric and comparing the induced judgments with human references.

Benchmark	Primary object	Core task or output	Criterion-validity audit	Criterion repair	Downstream validation
POPE (Li et al., 2023)	Model response	Object-existence hallucination	No	No	No
MMHal-Bench (Sun et al., 2023)	Model response	Hallucination and informativeness scores	No	No	No
HallusionBench (Guan et al., 2024)	Model response	Language-hallucination/visual-illusion diagnosis	No	No	No
AMBER (Wang et al., 2023)	Model response	Existence, attribute, and relation hallucinations	No	No	No
M-HalO (Cai et al., 2025)	Response token	Fine-grained hallucination localization	No	No	No
HalluScope (Jin et al., 2026)	Model response	Detection, type diagnosis, and corrective explanation	No	No	No
MLLM-as-a-Judge (Chen et al., 2024)	Multimodal judge	Response scoring and ranking	No	No	No
PerceptionRubrics (Wei et al., 2026)	Multimodal judge	Evaluation with atomic visual criteria	No	No	No
MRH-BENCH (ours)	Evaluation criterion	MRH type identification and rubric repair	Yes	Yes	Yes

Table 7: Source-task composition of MRH-BENCH. Modality descriptions summarize the visual inputs used from each source. T1, T2, and T3 denote the three benchmark tracks; five representative datasets summarize the long-tail sources.

Source dataset	Original task	Visual modality	T1	T2	T3	Total	Share (%)
MLLM-as-a-Judge (Chen et al., 2024)	Multimodal judge evaluation	Natural images, documents, and charts	416	406	1,451	2,273	26.8
Perception-Collection (Lee et al., 2024a)	Fine-grained rubric-based visual evaluation	Natural images	635	598	0	1,233	14.6
MM-RLHF (Zhang et al., 2025)	Multimodal preference alignment	Mixed visual inputs	331	322	524	1,177	13.9
GenExam (Wang et al., 2025)	Multidisciplinary text-to-image evaluation	Diagrams, maps, and plots	449	447	0	896	10.6
WaferMap-VQA (Xu & Lian, 2026)	Industrial wafer-map question answering	Industrial wafer maps	423	405	0	828	9.8
MIA-Bench (Qian et al., 2024)	Multimodal instruction following	Natural and mixed images	343	257	0	600	7.1
MMTutorBench (Yang et al., 2025)	Multimodal mathematical tutoring	Diagrams and documents	220	197	0	417	4.9
FACTS-Multimodal (Cheng et al., 2025)	Multimodal factuality evaluation	Images and documents	222	82	0	304	3.6
GRADE (Liu et al., 2026)	Image-editing reasoning	Edited images	127	105	0	232	2.7
Multi-Crit (Xiong et al., 2025)	Multi-criterion multimodal evaluation	Mixed visual inputs	112	82	0	194	2.3
InteractScience (Chen et al., 2025)	Interactive science understanding	Screenshots and plots	108	44	0	152	1.8
JudgeAnything-Image2Text (Pu et al., 2025)	Image-to-text evaluation	Mixed images	60	34	0	94	1.1
VL-RewardBench (Li et al., 2025)	Vision-language preference evaluation	Mixed visual inputs	14	13	25	52	0.6
UEval (Li et al., 2026a)	Rubric- and reference-based evaluation	Mixed visual inputs	12	8	0	20	0.2
Total			3,472	3,000	2,000	8,472	100.0

D BENCHMARK VERIFICATION PROTOCOL

Track 3 uses a separate annotation pass because its references concern response quality rather than benchmark-item validity. Absolute-score reliability uses ICC(2,3), a two-way random-effects, absolute-agreement intraclass correlation for the mean of the three ratings; this form matches the released reference, which is that mean rather than a single rating (Koo & Li, 2016). Ranking reliability uses Fleiss' κ over the ternary decision that each ranking induces on every response pair, $A \succ B$, $B \succ A$, or a tie (Landis & Koch, 1977). Across the 731 absolute-scoring cases (2,821 distinct responses) the three annotators reach ICC(2,3) $= 0.93$, 95% CI $[0.92, 0.94]$, indicating excellent reliability; across the 1,269 ranking cases (4,936 responses), Fleiss' $\kappa = 0.66$, 95% CI $[0.64, 0.69]$, indicating substantial agreement (Table 8). The released reference is the mean of the three scores; for rankings, a pairwise preference is retained only when at least two annotators share it, and the remaining pairs count as ties.

We validate each candidate in two independent layers. Kimi-K3 and Qwen3.8-Max first inspect the same evidence package without access to the construction rationale or to each other's decision. Table 9 reports their field-level results. Raw agreement stays above 89% for every field, but because both verifiers approve most candidates, expected agreement is already high and Cohen's κ stays between 0.24 and 0.36. We therefore screen on joint approval: a pair must pass both verifiers, which rejects 9.4% of pairs on average. Only jointly approved candidates proceed to expert review, which re-examines the same seven fields and approves substantially fewer of them (Table 10), so the model layer acts as a coarse filter while expert review carries the quality guarantee.

The low κ does not indicate a weak filter. Decomposing agreement into its positive and negative components (Cicchetti & Feinstein, 1990) shows why: positive specific agreement is 94.4–98.2% across fields, whereas negative specific agreement is only 24.6–41.4%. This is the marginal-imbalance paradox of κ (Feinstein & Cicchetti, 1990), which arises here because pass rates exceed 90% on every field. The two verifiers therefore agree closely on what to accept and disagree mainly on what to reject; since retention requires both to accept, the filter relies on the component where

Table 8: Pre-adjudication reliability of the three Track 3 annotators. Absolute scores are integers on a 0–5 scale. Counts denote distinct candidate responses rated by all three annotators. Unanimous agreement is the proportion of responses (absolute scoring) or cases (ranking) on which all three agree. ICC interval: analytic, F-based for $k = 3$ raters; κ interval: cluster bootstrap over cases, 2,000 resamples.

	Absolute scoring	Ranking
Cases	731	1,269
Distinct responses	2,821	4,936
Ratings per response	3	3
Reliability statistic	ICC(2,3)	Fleiss' κ
Estimate [95% CI]	0.93 [0.92, 0.94]	0.66 [0.64, 0.69]
Unanimous agreement (%)	40.1	53.6

Table 9: Independent model cross-validation before filtering. Percentages are over candidate pairs inspected by both verifiers. Agreement is raw; κ is Cohen's, chance-corrected. Joint pass requires both approvals.

Verification field	Kimi-K3 pass	Qwen3.8-Max pass	Agreement	κ	Joint pass
MRH validity	94.8	93.9	92.6	0.31	90.7
Task relevance	97.1	96.4	95.2	0.24	94.4
Type match	93.5	92.7	91.4	0.33	88.8
Evaluability	97.8	97.0	96.3	0.28	95.6
MRH resolution	92.9	91.8	90.5	0.33	87.6
Intent preservation	91.7	90.9	89.8	0.36	86.2
No new unsupported content	95.0	94.1	92.8	0.30	91.0
Macro average	94.7	93.8	92.7	0.30	90.6

they are consistent, and the weak negative agreement affects how many pairs are discarded rather than the quality of those retained.

After model filtering, three expert reviewers, all experienced practitioners in multimodal learning and NLP, independently receive the task image, prompt, task-authorized references, assigned MRH type, hallucinated criterion, and candidate repair, but not the generation rationale or one another's verdicts. They record a binary pass/fail decision on each of the seven fields used in the model layer (Table 10), reaching a macro Fleiss' κ of 0.83 with a minimum of 0.78 on intent preservation, the most subjective field. Disagreements are then discussed jointly until the three reviewers either reach consensus or agree to reject the item; only consensus-approved items enter the benchmark, and rejected items are regenerated or removed. Because the initial decisions precede any discussion, these κ values reflect independent judgment rather than convergence induced by deliberation.

E Taxonomy Boundaries and Annotation Rules

Annotation target. The six types in Table 1 identify the primary mechanism that makes a scoring requirement invalid. Each Track 1 instance is assigned exactly one MRH type label from T1–T6. The annotation unit is the smallest complete scoring requirement that can be evaluated independently, interpreted within the full rubric, task prompt, visual evidence, and authorized references. G/O/U/X specifies whether Evidence Support and Task Fit hold; T1–T6 specifies the mechanism of failure. A requirement is valid when it is supported by the authorized evidence and falls within the task scope. Reference wording and format become binding only when the task explicitly requires them.

Category boundaries. Type assignment depends on the source of the invalid requirement and the evidence needed to justify it, rather than on isolated keywords. Table 11 gives the defining condition and the distinguishing rule for each type. Task-scope additions and reference-derived constraints

Table 10: Three-expert agreement before adjudication. Pass rate is averaged over the three experts; unanimous agreement counts items where all three decide identically. Because Table 9 has two verifiers while this layer has three, their κ values are not directly comparable; the pass rates are.

Review field	Mean pass rate (%)	Unanimous agreement (%)	Fleiss' κ
MRH validity	80.5	92.4	0.84
Task relevance	81.7	94.1	0.87
Type match	79.7	90.8	0.81
Evaluability	81.2	95.0	0.89
MRH resolution	80.6	91.6	0.82
Intent preservation	78.3	88.9	0.78
No new unsupported content	80.6	92.1	0.83
Macro average	80.3	92.1	0.83

Table 11: Operational boundaries of the six MRH types. The distinguishing rules identify the primary failure mechanism of the targeted scoring requirement.

Type	Defining condition	Distinguishing rule
T1	**Task-Scope Overreach.** Adds an unrequested goal, deliverable, restriction, or prohibition.	The defect is an expansion of the task scope. A constraint specifically derived from a nonessential reference feature is T4; lack of evidence alone does not determine T1.
T2	**Unsupported Visual Claims.** Requires directly observable image content that is unsupported or contradicted, including objects, attributes, spatial relations, counts, OCR, and chart values.	The disputed claim concerns what the image directly depicts. A misread brand or date is T2; latent explanations are T3, and facts requiring additional records are T5.
T3	**Unwarranted Inference.** Converts limited observations into unjustified definite claims about intention, emotion, causality, future outcomes, or social meaning.	The defect is the inference from observations to a conclusion, not a direct visual reading (T2) or an external factual lookup (T5).
T4	**Reference Overconstraint.** Makes incidental reference wording, format, or details mandatory for receiving credit.	The constraint is traceable to a supplied reference or exemplar and is not required by the task. Explicitly requested reference features are valid constraints, not T4.
T5	**Unsupported External Facts.** Requires exact identities, attributes, or domain facts needing evidence beyond what the task provides or authorizes.	The missing support is external factual evidence. Directly readable content is assessed under T2; unavailable sensory information and provenance verification fall under T6.
T6	**Modality/Provenance Overreach.** Requires an unavailable sensory channel or unprovided evidence of origin, authenticity, or capture conditions.	The missing support is specifically sensory or provenance evidence. Reading displayed metadata is T2; verifying actual capture conditions from pixels alone is T6.

are distinguished by whether an incidental exemplar feature is made mandatory. Evidence-related failures are distinguished by direct visual perception, inference from observations, external factual knowledge, and unavailable sensory or provenance evidence.

Contrastive examples. The same subject matter can involve different mechanisms. Requiring a wrong reading of a clearly visible logo is T2, whereas requiring an unprovided exact manufacturing year is T5. Inferring a specific motive from a smile is T3, whereas requiring the person's unprovided name is T5. Requiring the music playing in a still-image scene or the camera settings from pixels alone is T6. In an image-description task, adding an unrequested action plan is T1; requiring exactly three sentences solely because the reference uses three is T4. These examples distinguish the evidential basis and origin of a requirement from its surface wording.

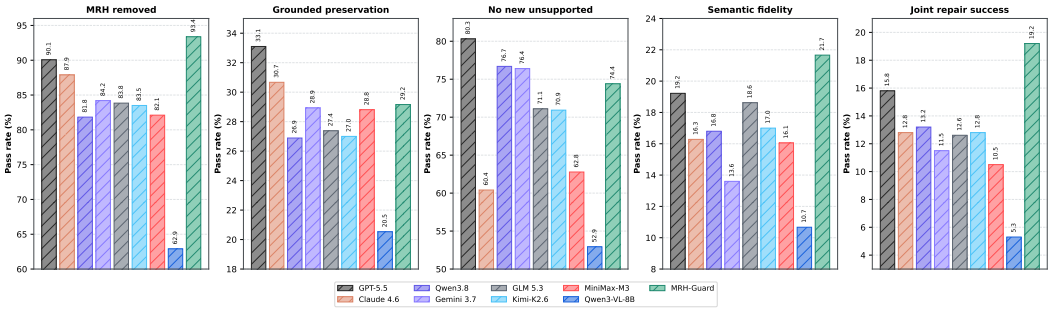


Figure 6: Track 2 criterion-level breakdown. For a common higher-is-better direction, R-MRH and New-Unsup are shown as their complementary pass rates, MRH removed and no new unsupported content. Each panel uses an independently truncated vertical scale to expose within-metric differences; values should therefore be compared within, not across, panels.

F Detailed Result Analysis

Figure 5 compares Track 1 Macro-F1 across methods. MRH-Guard improves Macro-F1 over the strongest prompted baseline by 0.144. Its Weighted-F1 improvement in Table 3 is 0.137, showing that the gain is not driven by class frequency.

Figure 6 decomposes repair quality into four marginal conditions and their conjunction. Individual MRH removal is relatively frequent, but preservation and semantic fidelity are substantially harder; consequently, joint RS is much lower than any single error-removal rate. MRH-Guard is strongest on MRH removal, semantic fidelity, and their conjunction, whereas GPT-5.5 is more conservative in preserving grounded content and avoiding new unsupported requirements.

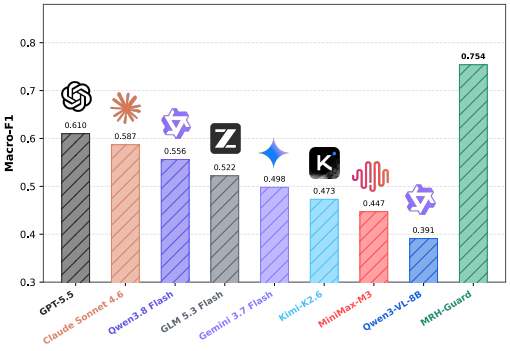


Figure 5: Track 1 Rubric hallucination type identification.

Figure 7 decomposes downstream alignment into absolute-scoring and ranking metrics. For absolute scoring, MRH-Guard reaches 62.55% Spearman correlation, 61.80% QWK, and 85.52% MAE agreement, outperforming the strongest prompted baseline in every panel. For ranking, it obtains the highest Kendall correlation at 49.37% and nearly matches GPT-5.5 in pairwise accuracy (77.15% versus 77.20%). The consistent gains across correlation, ordinal agreement, and error complement indicate that the repaired rubrics improve downstream decisions rather than only matching the surface form of the gold repair.

G Evaluation on Qwen3.5-9B

We additionally evaluate MRH-Guard on Qwen3.5-9B to test whether its gains extend beyond the original Qwen3-VL-8B backbone. Table 12 compares prompted inference, SFT, SFT+GRPO, and MRH-Guard across all three tracks. Qwen3-VL-8B results are reproduced from Tables 3–4 for comparison.

Results. On Qwen3.5-9B, MRH-Guard achieves the best results across all eight reported metrics. Relative to SFT+GRPO, it improves Macro-F1 by 1.5 points and RS by 1.50 percentage points, with gains of 0.0150 in Spearman's ρ and 0.0110 in PairAcc. It also outperforms MRH-Guard on Qwen3-VL-8B across all eight metrics, increasing Macro-F1 from 0.754 to 0.773 and RS from 19.20% to 20.60%. The consistent gains over training controls support the applicability of the framework to a second Qwen backbone.

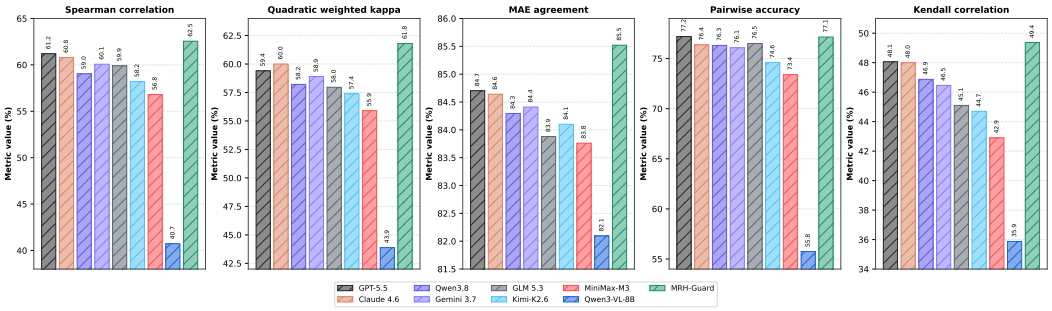


Figure 7: Track 3 downstream score alignment. Spearman's ρ, QWK, PairAcc, and Kendall's τ_b are expressed as percentages. MAE agreement is $100(1 - \mathrm{MAE}/5)$ under the 0–5 scoring scale, so all panels are higher-is-better. Each panel uses an independently truncated vertical scale; values should therefore be compared within, not across, panels.

Table 12: Cross-track comparison of Qwen3-VL-8B and Qwen3.5-9B. Track 2 values are percentages; the best results in each column are bold.

	Track 1	Track 2			Track 3			
Setting	Macro-F1 ↑	R-MRH (%) ↓	Sem-Fid (%) ↑	RS (%) ↑	ρ ↑	QWK ↑	MAE ↓	PairAcc ↑
Qwen3-VL-8B: main-table results								
Prompted	0.391	37.10	10.67	5.30	0.4072	0.4388	0.8953	0.5575
SFT	0.705	12.40	18.30	15.60	0.5862	0.5780	0.7930	0.7410
SFT+GRPO	0.741	8.90	20.10	17.60	0.6083	0.6020	0.7570	0.7584
MRH-GUARD	0.754	6.63	21.67	19.20	0.6255	0.6180	0.7240	0.7715
Qwen3.5-9B								
Prompted	0.417	24.80	15.10	9.40	0.5480	0.5410	0.8420	0.6810
SFT	0.726	10.70	19.70	17.00	0.6040	0.5900	0.7710	0.7560
SFT+GRPO	0.758	7.80	21.30	19.10	0.6270	0.6190	0.7390	0.7740
MRH-GUARD	**0.773**	**5.90**	**23.10**	**20.60**	**0.6420**	**0.6340**	**0.7080**	**0.7850**

H JUDGE VALIDATION AND RUBRIC CONTROLS

We assess judge–human agreement in Track 2, rubric controls in Track 3, and robustness to judge choice. Claude Fable 5 is the primary judge; the additional evaluation uses GPT-6 Astra (OpenAI, 2026c) with the same track-specific prompts and stored repair outputs.

H.1 TRACK 2: HUMAN CALIBRATION OF THE REPAIR AUDITOR

Setup. The calibration protocol samples 120 distinct source cases, 20 per MRH type, with source-stratified sampling and seed 42, independently of repair outcomes. It uses stored repairs from GPT-5.5, Claude Sonnet 4.6, Qwen3-VL-8B + SFT+GRPO, and MRH-GUARD: 480 outputs in total. Three experts independently assess each output using the same evidence and gold repair as the auditor, without seeing model identities or auditor decisions. They label residual MRH, grounded-content preservation, new unsupported content, and semantic fidelity following Appendix A. Human adjudication resolves disagreements. For both humans and the auditor, RS requires all four conditions to pass. Both judges are evaluated against the same adjudicated human labels on these 480 stored repairs; no additional human annotation is required.

Agreement measures. Table 13 reports pre-adjudication human agreement using Fleiss' κ, and judge–human agreement against the adjudicated reference using Cohen's κ and raw agreement. FalsePass is the fraction of human failures passed by the judge; FalseFail is the fraction of human passes rejected by the judge. Human-positive counts show the class balance.

Table 14 compares human RS with each judge on the same cases. Gap is judge RS minus human RS, measured in percentage points; a negative value indicates a lower aggregate pass rate under the judge. These descriptive subset comparisons are separate from the full-benchmark results in Table 2.

Table 13: Track 2 judge–human calibration. N counts repairs and N_H^+ counts human passes. κ_H: human Fleiss' κ; κ_{JH}: judge–human Cohen's κ. Agreement and error rates are percentages.

Condition	N	N_H^+	κ_H	κ_{JH}	Agreement	FalsePass	FalseFail
Claude Fable 5							
No residual MRH	480	436	0.805	0.853	97.50	11.36	1.61
Grounded-content preservation	480	152	0.893	0.850	93.54	4.27	11.18
No new unsupported content	480	353	0.895	0.846	93.96	10.24	4.53
Semantic fidelity	480	99	0.742	0.793	93.33	3.67	18.18
Joint repair success	480	81	0.855	0.796	94.38	3.01	18.52
GPT-6 Astra							
No residual MRH	480	436	0.805	0.818	96.88	13.64	2.06
Grounded-content preservation	480	152	0.893	0.879	94.79	3.35	9.21
No new unsupported content	480	353	0.895	0.842	93.75	9.45	5.10
Semantic fidelity	480	99	0.742	0.819	94.17	3.15	16.16
Joint repair success	480	81	0.855	0.833	95.42	2.26	16.05

Table 14: Human and judge RS on the Track 2 calibration subset. Each repair source has 120 cases. RS is in percent; Gap is judge RS minus human RS in percentage points (pp).

		Claude Fable 5		GPT-6 Astra	
Repair source	Human RS	Judge RS	Gap (pp)	Judge RS	Gap (pp)
GPT-5.5	17.50	15.83	-1.67	15.83	-1.67
Claude Sonnet 4.6	11.67	12.50	$+0.83$	9.17	-2.50
SFT+GRPO	18.33	17.50	-0.83	17.50	-0.83
MRH-GUARD	20.00	19.17	-0.83	21.67	$+1.67$

Results. Claude–human agreement ranges from 93.33% to 97.50%, with Cohen's κ between 0.793 and 0.853. Human RS and both judges' RS yield the same model ordering. Astra passes 77 of 480 repairs overall (16.04%). Judge-minus-human gaps range from -1.67 to $+0.83$ percentage points for Claude and from -2.50 to $+1.67$ for Astra. However, Claude rejects 18.52% of repairs that humans consider successful, so small aggregate RS gaps do not imply negligible instance-level errors. These descriptive subset comparisons do not establish statistical equivalence between judges and humans or significant differences between repair models.

H.2 TRACK 3: UNREPAIRED AND GOLD-REPAIR CONTROLS

Setup. The control protocol uses Claude Fable 5 on all 731 absolute-scoring and 1,269 ranking cases. It compares the original hallucinated rubric, repairs from the same four models used above, and the verified gold repair. Only the affected scoring unit changes; all other criteria, task evidence, candidate responses, human references, and judge settings remain fixed. The judge never sees human judgments or model identities. Gold repair is supplied only as the rubric in its own condition, not as extra evidence. Each condition uses three repeated judge calls with fixed repairs and the prompts in Appendix L, not three training seeds.

Metrics and comparisons. Table 15 reports Spearman's ρ, QWK, and MAE for scoring, and PairAcc and Kendall's τ_b for ranking, following Appendix A. Metrics are averaged over the three calls. Comparisons measure improvement over the original rubric and the remaining gap to gold repair; MAE uses the opposite direction because lower is better. The comparisons are descriptive. Original and gold rubrics are reference conditions, not guaranteed lower and upper bounds.

Results. All four model repairs improve all five metrics over the original hallucinated rubric. Under the fixed judge, MRH-GUARD increases Spearman's ρ from 0.5320 to 0.6255 and PairAcc from 0.7040 to 0.7715, while reducing MAE from 0.8820 to 0.7240. These gains show that rubric repair improves downstream agreement with human judgments, beyond changing the wording of the rubric. The verified gold repair performs best across all five metrics, indicating remaining room for improvement in model-generated repairs.

Table 15: Track 3 rubric controls under Claude Fable 5. N_S and N_R denote scoring and ranking case counts.

Rubric condition	N_S	N_R	$\rho \uparrow$	QWK $\uparrow$	MAE $\downarrow$	PairAcc $\uparrow$	$\tau_b \uparrow$
Original hallucinated rubric	731	1,269	0.5320	0.5060	0.8820	0.7040	0.3820
GPT-5.5 repair	731	1,269	0.6120	0.5940	0.7649	0.7720	0.4806
Claude Sonnet 4.6 repair	731	1,269	0.6080	0.6000	0.7680	0.7637	0.4800
SFT+GRPO repair	731	1,269	0.6083	0.6020	0.7570	0.7584	0.4740
MRH-GUARD repair	731	1,269	0.6255	0.6180	0.7240	0.7715	0.4937
Verified gold repair	731	1,269	0.6680	0.6510	0.6680	0.8010	0.5350

Table 16: Cross-judge comparison on matched subsets. Each method uses 600 Track 2 source cases and 220 scoring plus 380 ranking cases in Track 3. RS is a percentage; Track 3 metrics follow a three-run averaging protocol. Higher is better; bold marks the best results in each column.

	Claude Fable 5			GPT-6 Astra		
Repair source	RS (%) $\uparrow$	$\rho \uparrow$	PairAcc $\uparrow$	RS (%) $\uparrow$	$\rho \uparrow$	PairAcc $\uparrow$
GPT-5.5	16.00	0.6148	**0.7741**	16.33	0.6206	**0.7754**
Claude Sonnet 4.6	13.17	0.6069	0.7632	13.50	0.6121	0.7671
Qwen3-VL-8B + SFT+GRPO	17.83	0.6096	0.7593	17.67	0.6138	0.7627
MRH-GUARD	**19.17**	**0.6267**	0.7732	**19.00**	**0.6309**	0.7746

H.3 ROBUSTNESS TO JUDGE CHOICE

Setup. The cross-judge protocol uses 600 Track 2 source cases (100 per MRH type) and 600 Track 3 source cases (220 scoring and 380 ranking). Sampling uses seed 42 and source stratification within each type or evaluation setting, independently of repair outcomes. Both judges evaluate the same subsets and stored repairs from GPT-5.5, Claude Sonnet 4.6, Qwen3-VL-8B + SFT+GRPO, and MRH-GUARD. Task evidence, authorized references, prompts, and all candidate responses selected for evaluation remain fixed; repair-model identities are hidden. Track 2 uses one audit per repair. Track 3 uses three judge runs per candidate response for each judge and averages the metrics computed separately in each run. The human references remain fixed and are hidden from the Track 3 judges; repairs are not regenerated.

Comparison. Table 16 compares Track 2 RS and Track 3 Spearman's ρ and PairAcc. Both judges' metrics are computed on the same subsets, not copied from the full-benchmark tables. The comparison concerns method ordering and within-judge gaps, especially MRH-GUARD versus SFT+GRPO. The 480-output human calibration above remains a separate analysis.

I MECHANISTIC ABLATION METRICS

Table 17 complements the track-level ablation results with metrics tied directly to QAPO's diagnostic rewards. E-Acc and T-Acc measure unit-level Evidence Support and Task Fit accuracy, respectively, while Q-Acc requires the complete G/O/U/X quadrant to be correct. RS measures whether the resulting criterion satisfies all four Track 2 repair conditions. Together, these metrics test whether each reward improves its intended diagnostic decision and whether that improvement transfers to rubric repair.

All variants use the same backbone, data, rollout budget, and optimization steps. SFT Only removes Stage 2, while SFT+GRPO normalizes a single summed reward instead of normalizing

Table 17: Mechanistic ablation of the two-stage objective. E/T/Q-Acc denote Evidence Support, Task Fit, and joint-quadrant accuracy, respectively; RS is Track 2 repair success.

Variant	E-Acc $\uparrow$	T-Acc $\uparrow$	Q-Acc $\uparrow$	RS $\uparrow$
MRH-GUARD	**86.4**	**84.9**	**76.8**	19.20
SFT Only	80.7	79.2	68.5	15.60
SFT+GRPO	84.8	83.2	74.1	17.60
w/o r_E	79.9	84.1	71.6	16.93
w/o r_T	85.3	78.4	70.8	16.47
w/o r_Q	84.9	83.4	68.2	16.07
w/o r_P	85.8	84.3	75.6	17.93

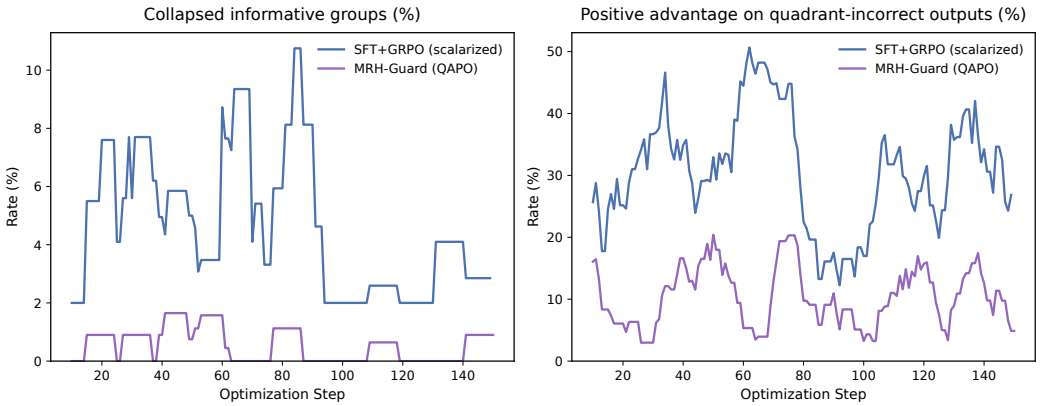


Figure 8: Credit-assignment dynamics of SFT+GRPO and QAPO over 150 optimization steps in a matched-budget diagnostic run. Each update uses eight prompts with four rollouts per prompt ($K = 4$; 32 rollouts total), and curves show 10-step moving averages. The main experiments instead use $K = 8$ with four prompts per update, preserving the same 32-rollout update budget. Left: the percentage of informative rollout groups whose channel-specific signal collapses under the optimization objective. Right: the percentage of quadrant-incorrect rollouts assigned positive advantage. Lower is better for both metrics.

each channel separately. The remaining variants remove one QAPO reward at a time. E-Acc, T-Acc, and Q-Acc provide the mechanism-level analysis; the main-text ablation reports representative end-task metrics from the three evaluation tracks.

Credit-assignment dynamics. Figure 8 directly tests the mechanism motivating QAPO. The figure uses a separate $K = 4$ diagnostic run to obtain more rollout groups per update at the same total rollout budget; it is used only for mechanism analysis, whereas every reported benchmark result uses the main $K = 8$ configuration. Compared with SFT+GRPO, QAPO consistently preserves more informative within-group reward variation and assigns positive advantage to substantially fewer quadrant-incorrect rollouts. Channel-wise normalization therefore retains axis-specific learning signals that can be suppressed by a single scalarized reward, while reducing updates that reinforce an incorrect Evidence–Task diagnosis.

J State-Transition Analysis of QAPO

We further analyze whether QAPO's axis-aware credit assignment produces the repair directions prescribed by Equation 2. For each scoring unit, we pair its pre-repair gold state with an independent post-repair G/O/U/X audit. The resulting transition matrices are row-normalized, so each row describes where content from one source state moves after repair. The desired destination is G for all invalid states, while G content should remain G.

Repair-transition fidelity. Figure 9 shows that QAPO increases the probability of every desired transition. G preservation rises from 92% to 94%, indicating that the additional optimization does not obtain its gains by aggressively rewriting already valid content. The improvements are larger for invalid states: O→G increases from 80% to 84%, U→G from 78% to 83%, and X→G from 67% to 73%. The increasing gain from G to X follows the geometry of the repair problem. O and U each require one invalid axis to be corrected while preserving the other, whereas X requires Evidence Support and Task Fit to be repaired jointly. QAPO therefore provides its largest benefit on the transition for which a scalar reward is most likely to hide a one-axis-only improvement.

Advantage assignment by error type. Figure 10 resolves the aggregate credit-assignment statistic in Figure 8 by quadrant decision. Under SFT+GRPO, 25–38% of off-diagonal outputs still receive positive advantage, despite predicting the wrong Evidence–Task state. QAPO reduces this

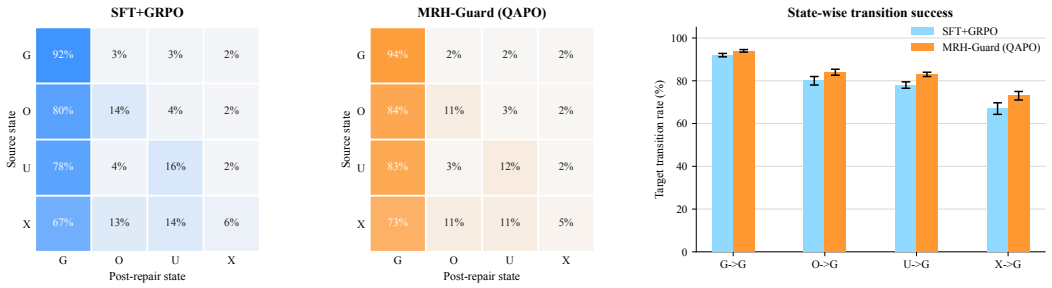


Figure 9: Post-repair quadrant transitions for SFT+GRPO and MRH-GUARD with QAPO. Rows denote the source state and columns the independently audited post-repair state; each row sums to 100%. The bar chart extracts the desired transition from each row: G→G measures preservation, whereas O→G, U→G, and X→G measure successful one- or two-axis correction.

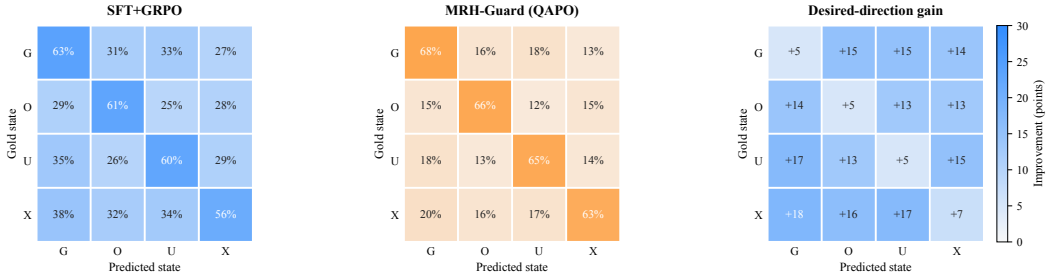


Figure 10: Positive-advantage rate conditioned on the gold and predicted quadrant. Diagonal cells are correct quadrant decisions; off-diagonal cells are diagnostic errors. The right panel reports the desired-direction gain of QAPO over SFT+GRPO: the QAPO-minus-SFT+GRPO difference on the diagonal and the reversed difference off the diagonal, so larger values consistently indicate better credit assignment.

range to 12–20%, corresponding to improvements of 13–18 points across individual error transitions. At the same time, the positive-advantage rate on correct diagonal decisions increases by 5–7 points. The effect is therefore asymmetric: QAPO modestly strengthens already correct rollouts but more strongly suppresses updates that would reinforce an incorrect repair direction.

These patterns explain the complementary roles of the four QAPO rewards. The axis rewards r_E and r_T preserve partial information about which boundary is correct, r_Q favors the exact G/O/U/X conjunction, and r_P prevents a gain on one axis from compensating for deterioration on the other. Their channel-wise normalization concentrates positive advantage on exact quadrant decisions and, in turn, raises direct transitions to G while maintaining G→G preservation. This state-level evidence complements the end-task ablation in Table 5 and the diagnostic accuracies in Table 17.

K TRAINING IMPLEMENTATION AND WEIGHT SELECTION

Token-level implementation. We mask all system, user, and multimodal input tokens and apply weighted cross-entropy only to the assistant target. Every supervised token is assigned to exactly one group. The opening and closing G/O/U/X tags receive w_{diag}; all tokens inside the `category` field receive w_{cat}; and all criterion tokens inside `fixed_rubric` receive w_{repair}. Criterion text copied inside the diagnostic field and all remaining structural tokens receive w_{copy}. Padding positions are excluded, and each example is normalized by its total active token weight as in Equation 3, so sequence length and padding do not change its overall scale.

Table 18: Sensitivity to the Stage 1 token weights. All variants keep $w_{\text{copy}} = 1$ and use the same training and evaluation protocol.

Setting	w_{diag}	w_{cat}	w_{repair}	Macro-F1 ↑	RS (%) ↑	ρ ↑	PairAcc ↑
Uniform	1	1	1	0.731	17.87	0.6124	0.7591
Moderate	3	3	2	0.747	18.70	0.6203	0.7670
Main	5	5	2	0.754	**19.20**	**0.6255**	**0.7715**
Strong	7	7	2	**0.757**	18.63	0.6218	0.7676

Main configuration. Only relative weights matter under this normalization, so we fix $w_{\text{copy}} = 1$. Our main configuration uses

$$(w_{\text{diag}}, w_{\text{cat}}, w_{\text{repair}}, w_{\text{copy}}) = (5, 5, 2, 1), \tag{8}$$

or equivalently $(\alpha, \beta) = (5, 2)$. This setting emphasizes the sparse diagnostic boundaries and category decision while retaining substantial supervision for positive repair. The same weights are used for every training instance and are not adjusted by MRH type or evaluation track.

Optimization setup. Stage 1 fine-tunes Qwen3-VL-8B-Instruct for three epochs with LoRA rank 64 and scaling factor 128. We use a learning rate of 1×10^{-5}, an effective batch size of 8, a maximum sequence length of 4,096, a warmup ratio of 0.05, and bfloat16 training. Stage 2 runs QAPO for one epoch with rollout-group size $K = 8$, training batch size 32, PPO mini-batch size 4, per-GPU micro-batch size 1, and actor learning rate 1×10^{-6}. Rollouts use temperature 0.7, top-p 0.8, and a maximum response length of 2,048. The clipping ratio, KL coefficient, and entropy coefficient are 0.2, 0.001, and 0.001, respectively. All runs use seed 42.

Data splits. The training corpus comprises three disjoint partitions: 5,000 instances for Stage 1 training, 599 instances for development, and 6,492 instances for Stage 2 reinforcement learning, totaling 12,091 instances. The partitions are obtained by a seeded shuffle with seed 42. No instance from the benchmark test tracks is used during training. The development set is used to monitor validation loss and select training hyperparameters. All reported results are computed on the held-out MRH-Bench tracks, which contain 3,472, 3,000, and 2,000 instances for Tracks 1–3, respectively.

Hardware and training cost. Training uses two NVIDIA H20 GPUs with 144 GB of memory per GPU, with PyTorch 2.10.0, Transformers 5.10.4, and CUDA 12.8. Stage 1 takes 8.0 wall-clock hours (16.0 GPU-hours), while Stage 2 takes 14.7 wall-clock hours (29.4 GPU-hours). Peak allocated memory is 76.9 GB per GPU. The complete training procedure therefore requires approximately 45.4 GPU-hours on our internal cluster; we do not report a monetary cost.

QAPO reward implementation. The main configuration fixes the channel mixture $(\lambda_E, \lambda_T, \lambda_Q, \lambda_P) = (0.2, 0.2, 0.4, 0.2)$ and uses $\epsilon = 1 \times 10^{-6}$ in Equation 6. These values are held constant across examples and training updates. We align predicted unit tags to the source units in their original order. A rollout is structurally invalid if it does not contain exactly one valid state for every source unit or if removing the state tags does not recover the source criterion. Such a rollout receives $r_E = r_T = r_Q = 0$, remains in the group, and participates in r_P and channel normalization with those values. If a reward channel has zero within-group variance, we set its normalized advantage to zero.

Clipped objective. Let π_{old} denote the behavior policy that generated a rollout and let π_{ref} be the frozen Stage 1 policy. For completion token t, we use

$$\rho_{i,t}(\theta) = \frac{\pi_\theta(z_{i,t} \mid x_i, z_{i,<t})}{\pi_{\text{old}}(z_{i,t} \mid x_i, z_{i,<t})}. \tag{9}$$

The sequence-level $A_{\text{QAPO}}(i)$ is broadcast to all unmasked completion tokens. We average the clipped surrogate, KL penalty to π_{ref}, and entropy bonus over those tokens and then over the rollout group. Prompt, image, and padding positions are excluded from all three terms. Thus, π_{old} supplies the importance ratio, whereas the independently frozen π_{ref} anchors the policy to the supervised checkpoint.

Reward scope. The four rewards deliberately assess only the deterministic, unit-level Evidence Support and Task Fit decisions: r_E and r_T provide partial credit, r_Q requires their exact conjunction, and r_P rewards jointly non-dominated rollouts rather than a compensatory scalar trade-off. T1–T6 classification and repaired text are not assigned separate Stage 2 rewards; they remain anchored by Stage 1 initialization and KL regularization. We evaluate any transfer from improved diagnosis to classification and repair using the held-out tracks rather than a learned repair checker.

Weight sensitivity. We compare the main setting with uniform weighting and two diagnostic-weight variants. All configurations use the same backbone, training split, example order, optimizer, number of updates, and QAPO configuration; only the Stage 1 token weights change. The main setting is selected using the held-out training validation split, without consulting MRH-Bench test labels. Table 18 reports end-task sensitivity across all three tracks.

L Evaluation Prompts

We report the exact task-facing templates used for the main experiments. At runtime, the bracketed placeholder is replaced with the task prompt, ordered image manifest, authorized references, and the track-specific rubric fields. Candidate responses and human judgments are excluded from repair prompts.

Track 1: Rubric hallucination type identification

System Prompt:

Role:
You are diagnosing Multimodal Rubric Hallucination (MRH).

Task:
The supplied rubric criterion is known to contain exactly one primary MRH mechanism. Classify it into exactly one of the following six types.

- **T1: Prompt/task boundary violation.** Adds a score-changing goal, restriction, required content, or prohibition outside the requested task.
- **T2: Visual perception boundary violation.** Requires a perceptual claim unsupported by or contradicted by the image.
- **T3: Inference boundary overclaim.** Turns limited evidence into a definite latent-state, causal, intentional, relational, or future-outcome claim.
- **T4: Reference/exemplar overconstraint.** Turns an incidental reference or exemplar detail into a mandatory scoring condition.
- **T5: External exact-fact violation.** Requires external factual or exact information unavailable from the supplied task evidence.
- **T6: Modality/provenance violation.** Requires information that the available modality or provenance cannot verify.

Rules:

- Select exactly one primary MRH type.
- Base the decision only on the task prompt, attached image(s), and authorized reference context.
- Return no explanation or Markdown outside the required JSON object.

Input:
You will receive the task prompt, ordered image manifest, authorized reference context, and the hallucinated rubric criterion.

STRICT OUTPUT (exact format):

```
{"primary_MRH_type":"T1_task_scope_unsupported"}
```

User Instruction:

```
{{MODEL_VISIBLE_EVIDENCE}}
```

Track 2: Hallucinated rubric repair

System Prompt:

Role:
You are repairing a hallucinated multimodal rubric criterion. Repair the criterion, not the task response.

Task:
Produce a complete replacement criterion that can be used directly for scoring. The repair must be supported by the available evidence and aligned with the task while preserving the grounded evaluative intent and all valid, task-relevant content.

Rules:

- Positively rewrite the complete affected scoring unit.
- Do not return a deletion-only fragment, a diagnosis of what is unverifiable, or a statement about what should not be required.
- Introduce no unsupported fact, restriction, format, example, or external knowledge.
- Use only the task prompt, attached image(s), and authorized reference context.
- Perform the analysis internally and return only the required JSON object.

Input:
You will receive the task prompt, ordered image manifest, authorized reference context, and the original hallucinated criterion.

STRICT OUTPUT (exact format):

```
{"repaired_criterion":"complete repaired criterion"}
```

User Instruction:

```
{{MODEL_VISIBLE_EVIDENCE}}
```

Track 2: Independent Audit

System Prompt:

Role:
You are a strict multimodal evaluator auditing a candidate rubric repair. Evaluate the repaired criterion, not the task response.

Task:
Compare the candidate repair against the task evidence, the original hallucinated criterion, and the complete verified gold repair. Valid paraphrases are allowed only when they preserve every material, score-changing requirement in the gold repair.

Rules:
Apply a conservative standard. Do not excuse an omission, weakening, genericization, scope change, or granularity change merely because the candidate retains a central idea. Return four independent binary decisions.

- **Residual MRH:** true iff any invalid requirement remains explicitly, implicitly, by presupposition, or through paraphrase.
- **Grounded-content preservation:** true iff all material grounded and task-relevant content is preserved as a complete, usable criterion.
- **New unsupported content:** true iff the candidate introduces any unsupported or out-of-scope score-changing content.
- **Semantic fidelity:** true iff evaluative target, scope, polarity, coverage, conditions, and granularity match the gold repair.

A deletion-only, fragmentary, generic, answer-only, or diagnostic-only output fails grounded-content preservation and semantic fidelity. Do not give the candidate the benefit of the doubt when material content is missing or ambiguous. The reason must identify any omission, residual requirement, or new unsupported content; if all decisions pass, state why the complete gold content is preserved.

Input:

You will receive the task prompt, attached image(s), authorized references, original hallucinated criterion, verified gold repair, and candidate repair. The gold repair is the canonical semantic specification, not additional task evidence.

STRICT OUTPUT (exact format):

```
{"residual_mrh":false,
 "grounded_content_preserved":true,
 "new_unsupported_content":false,
 "semantic_fidelity":true,
 "reason":"brief evidence-based explanation"}
```

User Instruction:

```
{{MODEL_VISIBLE_JUDGE_INPUT}}
```

Track 3: Three-Criterion Repair

System Prompt:

Role:
You are repairing a three-criterion multimodal rubric. Repair the rubric, not the candidate responses.

Task:
One scoring criterion contains a Multimodal Rubric Hallucination. Return the complete three-criterion rubric with the affected scoring unit positively rewritten so that it is supported by the available evidence and aligned with the task.

Rules:

- Preserve all grounded evaluative intent and all unaffected valid criteria.
- Do not use deletion-only or diagnostic wording.
- Introduce no new unsupported requirement.
- Candidate responses and human judgments are hidden.
- Use only the task prompt, attached image(s), and authorized reference context.
- Perform the analysis internally and return only the required JSON object.

Input:
You will receive the task prompt, ordered image manifest, authorized reference context, and the hallucinated three-criterion rubric.

STRICT OUTPUT (exact format):

```
{"repaired_rubric":{
  "1":"...",
  "2":"...",
  "3":"..."
}}
```

User Instruction:

```
{{MODEL_VISIBLE_EVIDENCE}}
```

Track 3: LLM-Human evaluation agreement assessment

System Prompt:

Role:
You are a fixed multimodal answer judge.

Task:
Evaluate one candidate response against the supplied task evidence and three-criterion rubric. Give each criterion a score from 0 to 5, where 0 means the response does not satisfy the criterion and 5 means it fully satisfies the criterion. Decimals are allowed. Then give a holistic overall score from 0 to 5 under the complete rubric; it need not be the arithmetic mean of the criterion scores.

Rules:

- Judge only what the task evidence and rubric establish.
- Do not infer hidden human ratings, dataset labels, model identity, provenance, or other candidate responses.
- Perform the evaluation internally.
- Return exactly one JSON object with no Markdown or additional text.

Input:
You will receive the task prompt, image evidence, repaired three-criterion rubric, and one candidate response.

STRICT OUTPUT (exact format):

```
{"criterion_scores":{"1":0,"2":0,"3":0},
 "overall_score":0}
```

User Instruction:

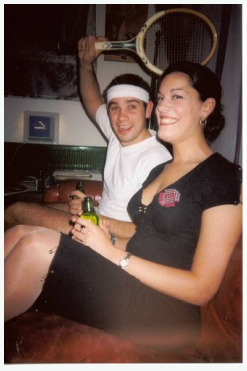
```
{{MODEL_VISIBLE_JUDGE_INPUT}}
```

M Case Study: What MRH-Bench Contains

Below, we present selected examples from MRH-Bench. Section M.1 includes two Track 1 examples of MRH type identification, and Section M.2 includes two Track 2 examples of rubric repair. Section M.3 presents four Track 3 examples of downstream score alignment, comprising two ranking cases and two absolute-scoring cases.

M.1 Track 1: MRH Type Identification

Example 1A — Unsupported Visual Detail



Input

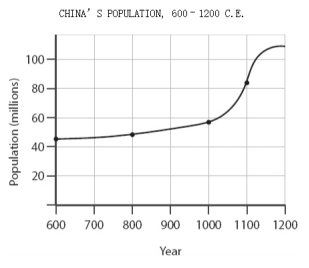
Task prompt. What is the woman wearing?

Rubric. *The response should identify the woman as wearing a black dress and may describe visible details such as its neckline and short sleeves. It must also state that the dress closes with a full-length silver zipper running down the front.*

Output. **T2: visual perception boundary violation.**

Interpretation. The dress and its visible styling are grounded, but the image does not reveal a full-length silver zipper. The required output is therefore a single T2 label.

Example 1B — Unrequested Output Constraint



Input

Task prompt. Draw a line graph of China's population from 600 to 1200 C.E., labeling the axes and showing the long plateau followed by the sharp late increase.

Rubric. *The graph should show the requested population trend with labeled axes and clearly marked data points. It must also include a companion table listing every plotted value.*

Output. **T1: prompt/task boundary violation.**

Interpretation. The graph-related requirements are requested by the prompt, whereas the companion table is an additional deliverable. This makes the target a T1 task-boundary violation rather than a visual-perception error.

M.2 Track 2: Rubric Repair

Example 2A — Unrequested Catering Requirements



Input

Task prompt. What is this, and can I make multiple variations of it for a party?

Rubric. *The response should identify the pictured food as an assortment of fried appetizers served with dipping sauce and propose several plausible variations for a party. It must also calculate exact ingredient quantities and a catering budget for exactly fifty guests.*

Output. *The response should identify the pictured food as an assortment of fried appetizers served with dipping sauce and propose several plausible party variations, such as different fillings, sauces, or preparation methods.*

Interpretation. The repair retains both requested outputs, identifying the food and suggesting variations, while removing the unrequested guest count, quantity calculation, and budget.

Example 2B — Contradicted Visual Attribute



Input

Task prompt. Examine the image and answer: "What shape is the sign?"

Rubric. *The response must identify the foreground sign as a regular octagon, noting its eight approximately equal sides and roughly* 135° *interior angles.*

Output. *The response should identify the foreground bus stop sign as circular, based on its continuous round outer boundary.*

Interpretation. The requested attribute is relevant, but the criterion contradicts the visible shape. The repair replaces the unsupported octagon description with the visually grounded circular boundary.

M.3 Track 3: Downstream Score Alignment

Ranking examples.

Example 3A — Unsupported Event Purpose



Input

Task prompt. What are they doing? First locate the relevant objects and then identify the activity.

Rubric.

1. The response should identify the people as flying kites.
2. It may describe the visible field, participants, and airborne kites while keeping the activity central.
3. The response must explain that the participants gathered specifically to rehearse for an upcoming kite-flying competition.

Repaired rubric.

1. The response should identify the people as flying kites.
2. It may describe the visible field, participants, and airborne kites while keeping the activity central.
3. Any statement about why the participants gathered should be explicitly framed as a tentative interpretation based on visible context.

Candidate responses.

A. *The image shows adults and children flying kites across a grassy field. Several participants are holding kite lines and looking upward, while multiple kites are visible above and around them; trees and a large outdoor stage-like structure form the background.*
B. *The people are flying kites in an open park. They are spread across the grass with their arms raised and lines extending toward the visible kites, indicating that kite flying is the central activity.*
C. *The group appears to be taking part in a community kite-flying gathering in the park. Adults and children are flying kites near an outdoor stage, although the image does not establish the event's exact purpose or organization.*
D. *The participants have gathered to rehearse for an upcoming kite-flying competition. They are practicing coordinated maneuvers and testing their kites in preparation for the formal event.*

Human annotation. $A \succ B \succ C \succ D$.

Interpretation. The repair model never sees these candidate responses. The hallucinated rubric would favor D's unsupported competition narrative, whereas the repaired rubric rewards grounded descriptions and places D last, matching the hidden human ranking.

Example 3B — Unverifiable Camera Settings



Input

Task prompt. Explain the photograph's aesthetic quality using visible properties of the image.

Rubric.

1. The response should give an overall aesthetic judgment and connect it to visible properties of the photograph.
2. It should discuss relevant features such as composition, the contrast between the cat's green eyes and dark fur, its upward gaze, and the focus on its face.
3. It must attribute the catchlights and background blur to a 50 mm lens used at $f/1.4$ and treat those settings as verified facts.

Repaired rubric.

1. The response should give an overall aesthetic judgment and connect it to visible properties of the photograph.
2. It should discuss relevant features such as composition, the contrast between the cat's green eyes and dark fur, its upward gaze, and the focus on its face.
3. It should explain how the visible catchlights, facial highlights, and softly blurred background emphasize the subject and create depth.

Candidate responses.

A. *The vivid green eyes contrast with the dark fur, while the upward gaze, warm facial highlights, sharp facial detail, and soft background create a striking portrait.*
B. *The centered cat and blurred background make the portrait visually appealing.*
C. *The image is appealing because it was shot with a 50 mm lens at $f/1.4$, which produced the catchlights and background blur.*

Human annotation. $A \succ B \succ C$.

Interpretation. This second ranking example uses the same fields. The repair replaces unverifiable capture settings with visible photographic effects, so the human ordering favors responses grounded in the image.

Absolute-scoring examples.

Example 3C — Counting Cube Facelets



Input

Task prompt. How many small square facelets are on the complete cube shown in the image?

Rubric.

1. The response should determine the total number of small square facelets on the complete cube. It must treat every visible face as a 3×3 grid before extrapolating to all six faces.
2. The identified grid dimensions and calculation should be traceable to the visible structure of the cube.
3. The final count should be stated clearly and supported by a concise calculation.

Repaired rubric.

1. The response should identify each face as a 7×7 grid and determine that the complete cube contains $6 \times 49 = 294$ facelets.
2. The identified grid dimensions and calculation should be traceable to the visible structure of the cube.
3. The final count should be stated clearly and supported by a concise calculation.

Candidate responses.

A. *The pictured puzzle is a 7×7 cube, so each face contains 49 small square facelets. Because a cube has six faces, the complete surface contains $6 \times 49 = 294$ facelets.*

B. *Treating the puzzle as a standard* 3×3 *cube gives nine facelets on each face. Because a cube has six faces, the complete cube therefore contains* $6 \times 9 = 54$ *facelets.*
C. *Each visible face appears to contain a* 7×7 *grid, or 49 facelets. Since three faces can be seen in the image, the cube has* $3 \times 49 = 147$ *facelets in total.*

Human annotation. **A:** $5/5$; **B:** $1/5$; **C:** $3/5$.

Interpretation. The hallucinated rubric forces an incorrect 3×3 reading despite the visible 7×7 grid. After repair, the complete six-face calculation receives full credit, the three-face calculation receives partial credit, and the erroneous 3×3 extrapolation receives a low score.

Example 3D — Identifying a Managed Habitat



Input

Task prompt. Are these giraffes in their natural habitat? Explain your answer using the image.

Rubric.

1. The response should answer whether the giraffes appear to be in their natural habitat and explain the conclusion.
2. The explanation should use visible evidence and express uncertainty where the setting cannot be established conclusively.
3. The response should be clear, concise, and internally consistent. It must also design a complete enrichment plan for housing the giraffes in a zoo.

Repaired rubric.

1. The response should answer whether the giraffes appear to be in their natural habitat and explain the conclusion.
2. The explanation should use visible evidence and express uncertainty where the setting cannot be established conclusively.
3. The response should be clear, concise, internally consistent, and focused on the question asked.

Candidate responses.

A. *The giraffes do not appear to be in their natural habitat. They are likely in a zoo or another managed wildlife enclosure, as indicated by the maintained ground, visitor walkway, railings, people, and constructed features surrounding the animals.*
B. *It is impossible to determine whether the giraffes are in their natural habitat because the image provides no reliable information about their surroundings or proximity to people.*
C. *The giraffes are probably in a managed enclosure rather than the wild. The people and constructed viewing area in the background suggest a zoo, although the exact facility cannot be identified from the image alone.*

Human annotation. **A:** $5/5$; **B:** $1/5$; **C:** $4/5$.

Interpretation. The enrichment-plan requirement is unrelated to the requested habitat judgment. Once it is removed, the detailed evidence-based answer receives full credit, the shorter but grounded answer remains strong, and the response that overlooks the visible enclosure cues receives a low score.

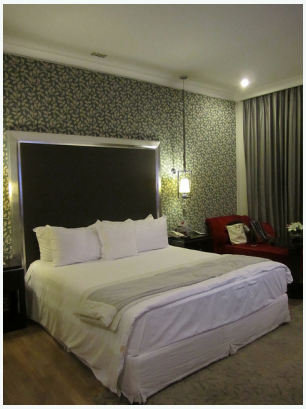
Table 19: How the three benchmark tracks expose progressively stronger targets.

Track	Model output	Hidden supervision
1	Single T1–T6 MRH type	Verified type annotation
2	Drop-in repaired criterion	Expert-verified positive repair and audit labels
3	Repaired criteria that drive a fixed judge	Human score or ranking on held-out responses

N Case Study: Training Data

The training file train.jsonl stores one instance per line. Each instance contains five main fields: image, prompt, original rubric, reference, and target text. The examples below show the structured G/O/U/X supervision used for training.

Example 1 — GX: Visible Interior Design and Unverifiable Authenticity



Prompt
Analyze the bedroom's interior design, including its furniture arrangement, color scheme, decorative features, aesthetics, and functionality.

Original rubric
The response should evaluate the room's furniture arrangement, color scheme, and decorative features, and determine whether the photograph is an unedited original.

Reference
The reference describes the bed, floral wallpaper, gray curtains, red chaise, decorative details, and lighting visible in the room.

Target text

[G] Evaluate the room's visible furniture arrangement, color scheme, and decorative features.

[X] Determine whether the photograph is an unedited original.

Category T6: modality/provenance violation

Fixed rubric
The response should evaluate the room's visible furniture arrangement, color scheme, decorative features, and lighting, and discuss how they affect its aesthetics and functionality.

Example 2 — GU: Correct Scene Analysis with a False Visual Detail



Prompt
Analyze the person's attire, the vehicle, and the surrounding environment, and discuss what historical setting the scene may evoke.

Original rubric
The response should interpret the scene's setting and possible historical significance. It must identify a motorcycle carrying the passengers.

Reference
The reference describes a formally dressed driver operating a horse-drawn carriage on a paved street.

Target text

[G] Interpret the setting and possible historical significance of the scene.

[U] Identify a motorcycle carrying the passengers.

Category T2: visual perception boundary violation

Fixed rubric
The response should interpret the setting and its historical character while correctly identifying the horse-drawn carriage and its formally dressed driver.

Example 3 — OX: Extra Creativity and an Unobservable Modality



Prompt
What is in the image?

Original rubric
The description should provide unique insights beyond stating "lobster on ice" and must state the lobster shell's exact smell.

Reference
The record contains source metadata but no additional content-bearing reference.

Target text

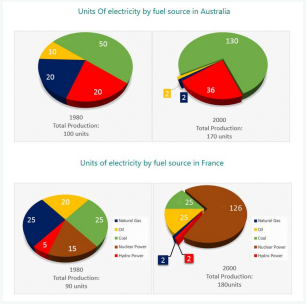
[O] Provide unique insights beyond stating "lobster on ice."

[X] State the lobster shell's exact smell from the image.

Category T6: modality/provenance violation

Fixed rubric
The response should describe the visible lobster resting on ice, including its mottled-brown shell, the nearby roses, and the small dish.

Example 4 — OU: Unrequested Style and a Contradicted Chart Value



Prompt
Summarize the electricity-production charts for Australia and France in 1980 and 2000, reporting the main features and relevant comparisons.

Original rubric
The summary should make fresh comparisons between the two countries and must report that Australia produced 180 total units in 2000.

Reference
The record contains source metadata but no additional content-bearing reference.

Target text

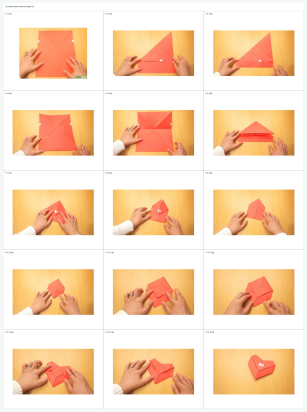
[O] Make fresh comparisons between Australia's and France's electricity-production trends.

[U] Report that Australia produced 180 total units in 2000.

Category T2: visual perception boundary violation

Fixed rubric
The response should summarize and compare the main trends, correctly reporting that Australia produced 170 units and France produced 180 units in 2000.

Example 5 — GO: Valid Step Alignment with an Exemplar Overconstraint



Prompt
How do I fold an origami heart? Show each step both visually and textually.

Original rubric
Each visual step must correspond to a text step and build on the previous one; the first step must use letter-size or A4 paper exactly as in the reference.

Reference
No separate textual reference is provided; the supplied image contains the visual folding sequence.

Target text

[G] Each visual step must correspond to a text step and progress cumulatively from the previous step.

[O] The first step must use letter-size or A4 paper exactly as in the reference.

Category T4: reference/exemplar overconstraint

Fixed rubric
Each visual step should correspond to a textual instruction and build on the previous step, beginning with a clearly shown sheet of paper of any suitable size.