

MM-VeriRubric: Verifiable Evidence-Grounded Rubrics for Multimodal Reward Models

Zhiyong Cao^{1,2,3,9*}, Bo Wang^{4,9*}, Jintao Li^{1,2,3,9*}, Shuxia Liu^{5,9}, Kairan Zhang⁹,
Aijia Wang⁶, Yi-Liang Wei^{7†}, Sicheng Zhou^{8,9†},

¹Institute of Automation, Chinese Academy of Sciences

²School of Artificial Intelligence, University of Chinese Academy of Sciences

³Zhongguancun Academy ⁴Institute of Hydrobiology, Chinese Academy of Sciences

⁵School of Astronautics, Northwestern Polytechnical University

⁶School of Education Science, Jiangsu Normal University

⁷School of Linguistic Sciences and Arts, Jiangsu Normal University

⁸Agent Efficiency Lab ⁹TraceLite AI

Correspondence: caozhiyong2024@ia.ac.cn, weiyiliang.2013@tsinghua.org.cn, zsicheng74@gmail.com

Abstract

Reward modeling is essential for multimodal alignment, yet most existing reward models only provide outcome reward, which is sparse and fails to capture human preferences. Rubrics can provide structured and fine grained evaluation signals, but building effective rubrics for multimodal tasks is challenging. In this work, we propose **MM-VeriRubric**, a framework for scalable construction of training data for multimodal rubric generation and rubric-based reward models. To better use visual evidence in images, we introduce task-aware visual clue extraction, which makes rubric generation more grounded in task-relevant visual information. MM-VeriRubric then generates evidence-grounded rubrics and constructs rubric-aligned rejected responses with controlled error dimensions, enabling the quality of each rubric to be verified through consistency checking. The verified data are used to train both a rubric generator and a rubric-based reward model. Experiments on multimodal reward model benchmarks and downstream Best of N tasks show that MM-VeriRubric improves preference judgment (+16.4% on VL-RewardBench) and provides more interpretable and diagnostic reward signals for multimodal alignment (+13.1% on MMMU-Pro using Best of 5 selection). Code is available at: <https://github.com/TraceLite-AI/MM-VeriRubric>.

1 Introduction

Vision language models (VLMs) are increasingly used as general purpose multimodal assistants for visual question answering, image reasoning, and document understanding (Liu et al., 2023). As these systems move toward real applications, multimodal alignment becomes essential because

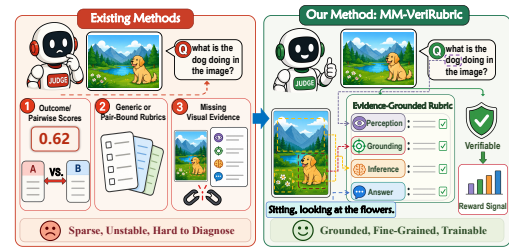


Figure 1: Comparison between existing outcome-level multimodal reward modeling and MM-VeriRubric, which builds evidence-grounded rubrics for fine-grained and verifiable reward signals.

their responses must be accurate, grounded in visual evidence, and aligned with human preferences (Ouyang et al., 2022; Rafailov et al., 2023; Yu et al., 2024). Reward models and model based judges provide key supervision for preference optimization (Wang et al., 2024). However, multimodal reward modeling is more difficult than text only reward modeling because response failures may arise from visual perception, evidence grounding, reasoning, answer formatting, or safety issues (Ruan et al., 2025). Existing outcome level rewards and pairwise preference methods often compress these heterogeneous failures into a single judgment, making it difficult to diagnose why a response fails and motivating more structured, fine grained reward supervision.

A single reward signal is sparse and lacks an attribution structure for decomposing response quality into different abilities (Liu et al., 2025). This limitation is especially problematic in multimodal tasks, where the same wrong answer may result from perception, grounding, reasoning, or answer formatting errors. Rubric reward addresses this

*Equal contribution.

†Corresponding author.

issue by making evaluation criteria explicit, allowing a response to be judged against concrete and checkable requirements rather than only through an overall preference (Gunjal et al., 2025). Such structured criteria provide fine grained and interpretable supervision, helping reward models learn how different error types affect response quality (Liu et al., 2025). However, multimodal rubrics cannot be treated as purely textual criteria, because a linguistically plausible but visually unanchored criterion may be difficult for a judge to execute consistently.

In multimodal reward modeling, rubrics must define response correctness while being tied to the visual evidence and operations required by the task. Existing rubric generation and rubric grounded reward modeling methods show the value of structured criteria, but they mainly rely on response pair comparison or rubric based filtering, rather than verifying instance level evidence support and stability under controlled error types (Liu et al., 2025; Kong et al., 2026). Therefore, the key data problem is not only to generate rubric text, but to construct task aware, evidence grounded, rubric aligned, and verifiable training examples.

To overcome this problem, we propose MM-VeriRubric, a framework for constructing verifiable rubric data for multimodal reward modeling. Instead of starting from response pairs alone, it first extracts visual clues that organize target entities, evidence anchors, required operations, query constraints, and anticipated failure modes. Based on these clues, it generates four dimensional rubrics over Perception, Grounding, Inference, and Answer, so that each criterion can be tied to visual evidence and a possible error source. It then constructs controlled rejected responses that deliberately corrupt specified dimensions while keeping the remaining aspects of the answer as stable as possible. Finally, a consistency verification step compares chosen and rejected responses under the rubric, filters unstable criteria, and converts verified samples into training data for rubric generation and rubric aware preference judgment.

Our contributions are summarized as follows.

- We propose MM-VeriRubric, a verifiable rubric data construction framework for multimodal reward modeling, which systematically produces supervision for training both a rubric generator and a rubric aware reward model.
- We design three core modules for evidence grounded rubric construction. It first extracts

visual clues and then generates four dimensional rubrics. Finally, it constructs controlled rejected responses and verifies whether the rubric can identify the intended error dimension.

- We build a multimodal rubric dataset for training rubric generation and rubric aware preference judgment. Experiments on multimodal reward model benchmarks and downstream Best of N tasks show that MM-VeriRubric improves preference judgment and provides more fine grained diagnostic reward signals.

2 Related Works

2.1 Multimodal Reward Modeling

Multimodal reward models extend the reward modeling paradigm from text-only RLHF and preference optimization to vision-language or broader multimodal settings, where they provide reward or preference signals for multimodal RLHF, direct preference optimization, response selection, and VLM-as-a-judge evaluation (Ouyang et al., 2022; Rafailov et al., 2023; Sun et al., 2024; Zhang et al., 2025c,a). Existing approaches mainly fall into three paradigms: learning reward models from annotated preference data (Sun et al., 2024; Yu et al., 2024, 2025; Zhang et al., 2025c), deriving preference signals from strong VLM or MLLM judges (Xiong et al., 2025), and developing specialized reward models for benchmark-specific evaluation (Li et al., 2025; Ruan et al., 2025; Tu et al., 2025; Wang et al., 2025a; Zhang et al., 2025b). However, most multimodal reward models still provide limited fine-grained evidence for why a response is preferred, which motivates more interpretable and controllable forms of reward supervision (Liu et al., 2026b).

2.2 Rubric-Based Reward Modeling

Rubric-based reward modeling makes reward signals more explicit than holistic scores or pairwise preferences by decomposing evaluation into natural-language criteria, providing fine-grained supervision (Gunjal et al., 2025; Viswanathan et al., 2026; Liu et al., 2025). Recent work studies rubric generation for prompts or response pairs (Liu et al., 2025, 2026a), rubric-guided reward models that produce criterion-level judgments (Gunjal et al., 2025; Viswanathan et al., 2026; Liu et al., 2026b), and rubric-grounded preference synthesis or alignment rewards (Kong et al., 2026; Yu et al., 2026).

However, in multimodal settings, rubrics must be tied to visual evidence rather than inferred only from textual instructions, so that judges can reliably localize errors in perception, grounding, reasoning, or final answers. (Chen et al., 2026; Liu et al., 2026b; Yu et al., 2026). Moreover, pair-conditioned or generalized rubrics are not well suited for scalable training-data synthesis for rubric-aware reward models. In contrast, MM-VeriRubric builds evidence-grounded multimodal rubric data with controlled negative samples, enabling reward models to learn structured and verifiable evaluation signals.

3 Preliminaries

Rubrics. Given a multimodal input $x = (I, q)$, where I denotes an image or a set of images and q denotes a text question or instruction, a model produces a candidate response y . We denote a reference answer or a high quality response as y^+ . For each input x , a rubric is a set of natural language evaluation criteria:

$$\mathcal{C}(x) = \{c_1, c_2, \dots, c_K\} \quad (1)$$

Each criterion c_k specifies one checkable requirement that a response should satisfy, such as identifying a visual entity, grounding a claim to an evidence region, performing a reasoning operation, or following the required answer format. In multimodal tasks, a useful criterion should not depend only on textual semantics, but should also be connected to visual evidence, task goals, and possible sources of error.

Rubrics as Reward. A standard reward model maps an input and a response to a scalar reward:

$$R_\theta(x, y) \in \mathbb{R} \quad (2)$$

Given a preference pair (y^+, y^-) , preference learning usually encourages the reward of the preferred response to be higher:

$$R_\theta(x, y^+) > R_\theta(x, y^-) \quad (3)$$

Rubrics as Reward instead conditions the reward model on a rubric $\mathcal{C}$, leading to a rubric-conditioned reward $R_\theta(x, y, \mathcal{C})$. For each criterion c_k , the model or judge can produce a criterion-level judgment

$$s_k = g_\theta(x, y, c_k), \quad s_k \in \{0, 1\} \quad (4)$$

or more generally $s_k \in [0, 1]$. The final reward can then be aggregated from criterion-level signals:

$$R_\theta(x, y, \mathcal{C}) = A(s_1, s_2, \dots, s_K) \quad (5)$$

where $A(\cdot)$ can be a sum, an average, or a learned aggregation function. Therefore, rubric reward not only compares response quality but also identifies which criteria or error dimensions a response fails to satisfy, which provides the basis for constructing verifiable rubrics and rubric aligned preference data.

4 Methodology

MM-VeriRubric is a data construction framework for building verifiable multimodal rubrics and rubric aligned preference data from image text question answering samples (Fig.2). Given an input $x = (I, q)$ and a reference answer y^+ , the framework produces two forms of supervision: rubric generation data $(I, q, y^+, v) \rightarrow \mathcal{C}$, and rubric aware reward modeling data $(I, q, \mathcal{C}, y^+, y^-)$. The overall process can be written as

$$\begin{aligned} v &= \phi_{\text{vcp}}(I, q), \\ \mathcal{C} &= \phi_{\text{rub}}(I, q, y^+, v), \\ \mathcal{Y}^- &= \phi_{\text{rej}}(I, q, y^+, \mathcal{C}) \end{aligned} \quad (6)$$

where v is the visual clue profile, $\mathcal{C}$ is the generated rubric, and $\mathcal{Y}^-$ is a set of controlled rejected responses.

The first step extracts task-aware visual clues from I and q , including target entities, evidence anchors, required operations, query constraints, and anticipated failure modes. The second step generates evidence grounded rubrics conditioned on I, q, y^+ , and v , with criteria organized into Perception, Grounding, Inference, and Answer dimensions. The third step constructs rejected responses that intentionally violate specified rubric dimensions while preserving the other aspects of the reference answer as much as possible. Finally, MM-VeriRubric verifies whether the rubric can distinguish y^+ from each rejected response under rubric conditioned scoring, filters unstable criteria, and converts the verified samples into training data for both the rubric generator and the rubric aware reward model.

4.1 Task-Aware Visual Clue Extraction

The first step of MM-VeriRubric extracts a task aware visual clue profile from the multimodal input

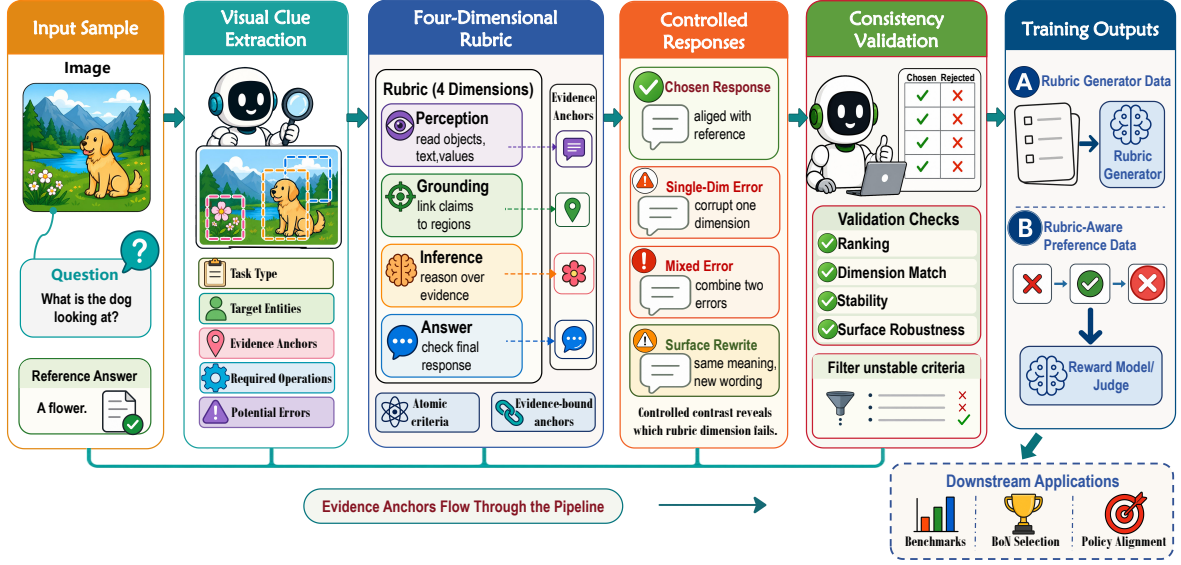


Figure 2: Overview of MM-VeriRubric. The framework extracts task-aware visual clues, generates evidence-grounded four-dimensional rubrics, constructs controlled rejected responses, and applies consistency verification to produce training data for rubric generation and rubric-aware reward modeling.

$x = (I, q)$, which summarizes the visual evidence and operations required to answer the question. We denote this step as

$$v = \phi_{\text{vcp}}(I, q) \quad (7)$$

where ϕ_{vcp} is a visual clue extractor implemented by a teacher VLM. The visual clue profile is represented as

$$v = \{E, A, O, T, F\} \quad (8)$$

where E denotes target entities, A denotes evidence anchors, O denotes required operations, T denotes query constraints and answer type, and F denotes anticipated failure modes.

The target entities E record the objects, text spans, regions, or visual elements that are relevant to the question. The evidence anchors A specify where the relevant information appears in the image, using forms such as object references, OCR spans, region descriptions, or bounding boxes. The required operations O describe the actions needed to answer the question, such as identifying, counting, comparing, calculating, locating, or reasoning over visual evidence. The query constraints and answer type T capture format requirements, units, option choices, and other constraints imposed by the question. The anticipated failure modes F describe possible errors that a model may make, such

as OCR misreading, entity confusion, wrong spatial relation, chart value misreading, or invalid reasoning.

The task aware visual clue profile is not intended to replace the original image. Instead, it serves as a structured intermediate representation that makes salient visual evidence and task requirements explicit. In the following rubric generation step, MM-VeriRubric conditions on I, q, y^+ , and v , so that the generated criteria are more concrete and more closely grounded in the image.

4.2 Evidence-Grounded Rubric Generation

After obtaining the visual clue profile, MM-VeriRubric generates an evaluation rubric that is grounded in the image evidence and aligned with the question goal. We formulate the rubric generation step as

$$\mathcal{C} = \phi_{\text{rub}}(I, q, y^+, v) \quad (9)$$

where y^+ is the reference answer, v is the visual clue profile, and ϕ_{rub} denotes a teacher model or a trained rubric generator. The output rubric is a set of criteria

$$\mathcal{C} = \{c_1, c_2, \dots, c_K\} \quad (10)$$

Each criterion is represented as

$$c_k = (d_k, r_k, a_k, t_k) \quad (11)$$

where d_k is the evaluation dimension, r_k is a natural language requirement, a_k is an evidence anchor, and t_k is a failure mode tag or scoring rule.

We organize criteria into four dimensions: Perception, Grounding, Inference, Answer.

Perception criteria check whether the response correctly reads or identifies visual elements. Grounding criteria check whether claims are linked to the correct visual evidence. Inference criteria check whether the response performs valid reasoning over the perceived and grounded evidence. Answer criteria check whether the final response satisfies the required content, format, unit, option, or safety constraint.

Evidence grounding is a key requirement in rubric generation. For Perception, Grounding, and Inference criteria, the evidence anchor a_k should refer to a concrete visual clue, such as an OCR span, an object reference, a region description, or a bounding box. The number of criteria is not fixed across samples. Instead, MM-VeriRubric allocates criteria according to the required operations and anticipated failure modes in v , so samples involving OCR, chart reading, spatial binding, or multi step reasoning can receive denser coverage. The resulting rubric serves not only as an evaluation specification, but also as the basis for rejected response construction and consistency verification, so each criterion is encouraged to be atomic, verifiable, and non redundant.

4.3 Rubric-Aligned Rejected Response Construction

After generating the rubric, MM-VeriRubric constructs controlled rejected responses to form verifiable preference pairs. For each sample, we treat the reference answer as the chosen response:

$$y^+ = y_{\text{ref}} \quad (12)$$

We consider four controllable error dimensions, for each target dimension $d \in \mathcal{D}$, we generate a rejected response by

$$y_d^- = \phi_{\text{rej}}(I, q, y^+, \mathcal{C}, d) \quad (13)$$

where ϕ_{rej} is a rejected response generator implemented by a teacher model.

The generated response y_d^- is required to violate criteria associated with the target dimension d , while keeping the other dimensions in $\mathcal{D} \setminus \{d\}$ as unchanged as possible. A Perception corruption changes what the response reads or identifies from

the image. A Grounding corruption preserves the perceived information but links it to a wrong region, entity, or evidence source. An Inference corruption keeps the relevant evidence but introduces an invalid reasoning step. An Answer corruption keeps the supporting process mostly intact but changes the final answer, format, unit, or option.

Beyond single-dimension corruptions, MM-VeriRubric can also construct mixed corruptions y_{d_i, d_j}^- to simulate more complex failure modes involving two dimensions. We define the single-dimension and mixed rejected response sets as

$$\begin{aligned} \mathcal{Y}_{\text{single}}^- &= \{y_d^- \mid d \in \mathcal{D}\}, \\ \mathcal{Y}_{\text{mix}}^- &= \{y_{d_i, d_j}^- \mid d_i, d_j \in \mathcal{D}, d_i \neq d_j\} \end{aligned} \quad (14)$$

and let

$$\mathcal{Y}^- = \mathcal{Y}_{\text{single}}^- \cup \mathcal{Y}_{\text{mix}}^- \quad (15)$$

In addition, we construct a semantic preserving surface rewrite y^0 , which changes only the wording of y^+ while preserving its meaning. These responses allow the framework to test both sensitivity to true errors and robustness to harmless surface variations. The resulting preference set is

$$\mathcal{P}(x) = \{(y^+, y^-) \mid y^- \in \mathcal{Y}^-\} \quad (16)$$

These controlled preference pairs are not directly kept as final training data, but are passed to the consistency verification step to test whether the rubric can recognize the injected error dimension.

4.4 Consistency Verification and Model Training Data Construction

After constructing controlled rejected responses, MM-VeriRubric performs consistency verification to check whether the generated rubric is executable, stable, and aligned with the injected error dimensions. For a response y and a criterion c_k , we define criterion level scoring as

$$s_k(y) = g_\theta(I, q, y, c_k), \quad s_k(y) \in \{0, 1\} \quad (17)$$

where g_θ denotes a rubric conditioned judge. The overall rubric score of y is computed by

$$S(y; \mathcal{C}) = \sum_{k=1}^K s_k(y) \quad (18)$$

For each rejected response $y^- \in \mathcal{Y}^-$, the rubric should prefer the reference response:

$$S(y^+; \mathcal{C}) > S(y^-; \mathcal{C}) \quad (19)$$

We further define the criteria associated with dimension d as

$$\mathcal{C}_d = \{c_k \in \mathcal{C} \mid d_k = d\} \quad (20)$$

For a single-dimension rejected response y_d^- , the score drop on dimension d' is

$$\Delta_{d'}(y_d^-) = \sum_{c_k \in \mathcal{C}_{d'}} (s_k(y^+) - s_k(y_d^-)). \quad (21)$$

A valid rejected response should mainly reduce the score of the target dimension, which can be checked by

$$\Delta_d(y_d^-) \geq \Delta_{d'}(y_d^-), \quad \forall d' \neq d. \quad (22)$$

At the same time, the non target dimensions should remain stable, so large unexpected drops on $\mathcal{C}_{d'}$ indicate that the corruption has changed more than the intended error type. For the semantic preserving surface rewrite y^0 , the rubric should give a similar score to the reference answer:

$$|S(y^+; \mathcal{C}) - S(y^0; \mathcal{C})| \leq \epsilon. \quad (23)$$

Criteria that fail on y^+ , over penalize y^0 , or cannot distinguish the injected error are removed or marked as unstable.

The verified samples are then converted into two types of training data. We denote the filtered rubric as

$$\tilde{\mathcal{C}} = \text{Filter}(\mathcal{C}, y^+, \mathcal{Y}^-, y^0). \quad (24)$$

For rubric generation, each sample is represented as

$$\mathcal{D}_{\text{gen}} = \{(I, q, y^+, v) \rightarrow \tilde{\mathcal{C}}\}, \quad (25)$$

and the generator is trained with the autoregressive objective

$$\mathcal{L}_{\text{gen}} = -\log p_{\theta}(\tilde{\mathcal{C}} \mid I, q, y^+, v). \quad (26)$$

For rubric aware reward modeling, each verified preference pair is converted into a supervised judge generation sample:

$$\mathcal{D}_{\text{rm}} = \{(I, q, \tilde{\mathcal{C}}, y_a, y_b) \rightarrow z\}, \quad (27)$$

where y_a and y_b are the two candidate responses in randomized order, and z contains criterion level judgments and the final preference decision. The rubric aware reward model is also trained with an autoregressive SFT objective:

$$\mathcal{L}_{\text{rm}} = -\log p_{\theta}(z \mid I, q, \tilde{\mathcal{C}}, y_a, y_b). \quad (28)$$

Through this verification and data construction process, MM-VeriRubric turns teacher generated visual clues, evidence grounded rubrics, and controlled preference pairs into trainable multimodal reward modeling data with explicit diagnostic structure.

5 Experiments

5.1 Experimental Setup

Training Data. We construct MM-VeriRubric training data from a mixture of image text instruction and reasoning datasets, including ScienceQA(Saikh et al., 2022), MathV360K(Shi et al., 2024), VLGuard(Zong et al., 2024), ShareGPT4V(Chen et al., 2024b), ALLaVA(Chen et al., 2024a), and MANTIS(Jiang et al., 2024). These sources cover general visual question answering, mathematical reasoning, safety related queries, open ended instruction following, and multi image understanding. We use the pipeline described in Section 4 to generate visual clue profiles, evidence grounded rubrics, controlled rejected responses, and verified preference pairs. The final training set contains 15,000 samples for the rubric generator and 10,000 samples for the rubric aware reward model. All evaluation benchmarks are kept as held out test sets and are not used in rubric synthesis or reward model training.

Models and Training. We train two types of models. The first model is a rubric generator, which takes an image, a question, a reference answer, and the visual clue profile as input, and outputs an evidence grounded rubric. The second model is a rubric aware reward model, which compares two candidate responses under the generated rubric and predicts the better response together with criterion level judgments. Both models are trained with supervised fine tuning using the LLaMA-Factory multimodal conversation format. We report two model scales, MM-VeriRubric-4B and MM-VeriRubric-8B, to study whether the proposed data construction framework remains effective across different backbone sizes. During inference, the rubric generator first produces an instance specific rubric for each test sample, and the reward model then performs rubric conditioned pairwise comparison.

Evaluation Benchmarks. We evaluate reward modeling performance on three held out multimodal reward model benchmarks: VL-RewardBench(Li et al., 2025), MM-

Table 1: **Evaluation Results of Multimodal Reward Modeling.** We compare MM-VeriRubric with proprietary models, open-source models, and open-source reward models. All benchmarks are evaluated by pairwise accuracy.

Model / Method	VL-RewardBench	MM-RewardBench	MMRB2-Reasoning	Avg.
<i>Proprietary Models without Rubrics</i>				
GPT-4o	65.8	71.5	51.9	63.1
Claude-3.5-Sonnet	55.3	72.0	—	—
Gemini-2.5-Pro	71.2	73.3	66.6	70.4
<i>Open-Source Models</i>				
Qwen3-VL-4B	52.2	59.8	50.8	54.3
Qwen3-VL-8B	54.1	60.1	51.2	55.1
<i>Open-Source Reward Models</i>				
Skywork-VL-Reward-7B	61.2	67.0	53.4	60.5
IXC-2.5-Reward-7B	66.6	68.8	52.3	62.6
Omni-RRM-7B (sft)	60.4	61.0	51.5	57.6
<i>Ours</i>				
MM-VeriRubric-4B	69.3	67.7	52.6	63.2
MM-VeriRubric-8B	70.5	70.2	53.8	64.8

Table 2: **Ablation Studies of MM-VeriRubric-8B.** We evaluate the contribution of key rubric construction components on VL-RewardBench. All results are reported by pairwise accuracy.

Variant	VL-RewardBench
Full MM-VeriRubric	70.5
w/o visual clue extraction	67.4
w/o rubric filter	62.3
reference answer only	54.7

Table 3: **Downstream Best-of-N Evaluation.** We compare single-sample generation with Best-of-N selection using MM-VeriRubric-8B. Results are reported using each benchmark’s official metric.

Dataset	Single	BoN	Gain
HallusionBench	63.4	65.8	+2.4
MathVista	64.6	67.9	+3.3
MMMU-Pro	45.3	58.4	+13.1

RewardBench(Yasunaga et al., 2025), and MMRB2-Reasoning(Hu et al., 2025). To further test whether the reward model can serve as an inference time selection signal, we also conduct Best of N experiments on HallusionBench(Guan et al., 2024), MathVista(Lu et al., 2024), and MMMU-Pro(Yue et al., 2025).

Baselines and Metrics. We compare MM-VeriRubric with four groups of baselines. The first group includes proprietary multimodal judges without rubrics, such as GPT-4o(Hurst et al., 2024), Claude-3.5-Sonnet, and Gemini-2.5-Pro(Comanici et al., 2025). The second group includes open source VLM judges, such as Qwen3-VL-4B and Qwen3-VL-8B(Bai et al., 2025). The third group

includes existing open source reward models, such as Skywork-VL-Reward-7B(Wang et al., 2025b), IXC-2.5-Reward-7B(Zang et al., 2025), and Omni-RRM-7B. For reward model benchmarks, we report pairwise accuracy on each benchmark and the average accuracy across benchmarks. For downstream Best of N evaluation, we report each benchmark’s official task metric after selecting the best response with the reward model.

5.2 Evaluation Results of Multimodal Reward Modeling

Table 1 reports pairwise accuracy on three multimodal reward modeling benchmarks. MM-VeriRubric achieves strong performance among open source models and reward models. The 8B model obtains an average accuracy of 64.8, outperforming Qwen3-VL-8B, Omni-RRM-7B, Skywork-VL-Reward-7B, and IXC-2.5-Reward-7B by 9.7, 7.2, 4.3, and 2.2 points, respectively. This shows that evidence grounded rubrics and verified preference pairs provide effective supervision for multimodal reward modeling.

On VL-RewardBench, MM-VeriRubric-8B reaches 70.5, achieving the best result among open source methods and approaching Gemini-2.5-Pro. On MM-RewardBench, it also outperforms all open source reward model baselines, indicating that rubric aware comparison helps the model judge responses through explicit criteria rather than only holistic preference signals. On MMRB2-Reasoning, all open source methods remain far behind Gemini-2.5-Pro, showing that reasoning intensive reward modeling is still challenging. Nevertheless, MM-VeriRubric-8B achieves the best open

source result on this benchmark.

Comparing the two model scales, MM-VeriRubric-8B consistently improves over MM-VeriRubric-4B across all benchmarks. This suggests that the proposed data construction framework scales with model capacity and can improve multimodal preference judgment without relying on proprietary judges at inference time.

5.3 Ablation Studies

Table 2 shows the ablation results on VL-RewardBench. Removing visual clue extraction decreases accuracy from 70.5 to 67.4, indicating that task aware visual evidence helps generate more grounded and useful rubrics. Removing the rubric filter causes a larger drop to 62.3, showing that consistency verification is important for filtering unstable criteria and invalid preference pairs.

The reference answer only variant performs the worst, with an accuracy of 54.7. This suggests that rubrics cannot be reliably constructed from the final answer alone, because the model loses access to the question context, visual evidence, and task specific failure modes. Overall, these results show that both evidence grounding and verification are necessary for effective rubric based multimodal reward modeling.

5.4 Downstream Best-of-N Evaluation

Table 3 and Fig.3 reports downstream Best of N results using MM-VeriRubric-8B as the selection model. Compared with single sample generation, Best of N selection improves performance on all three benchmarks, showing that MM-VeriRubric can provide useful reward signals for inference time response selection.

The gains are 2.4 points on HallusionBench, 3.3 points on MathVista, and 13.1 points on MMMU-Pro. The largest improvement appears on MMMU-Pro, suggesting that rubric aware selection is especially helpful for challenging multi domain reasoning tasks where candidate responses differ in perception, grounding, reasoning, and final answer quality. These results indicate that MM-VeriRubric is not only effective for reward model benchmarks, but can also improve downstream multimodal reasoning through response selection.

6 Conclusion

Multimodal reward modeling requires fine grained, interpretable, and diagnostic evaluation signals, because outcome level rewards often cannot explain

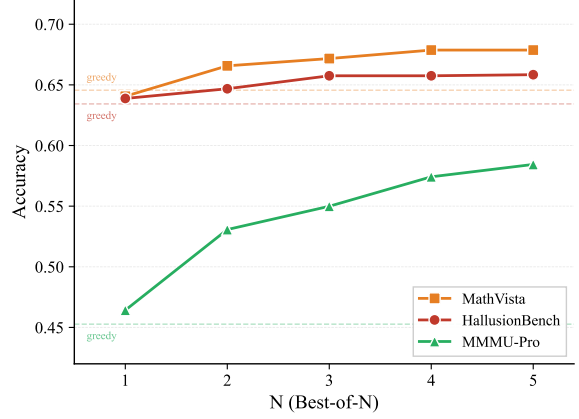


Figure 3: Best-of-N Select.

whether an error comes from visual perception, evidence grounding, reasoning, or final answer formulation. In this paper, we introduced MM-VeriRubric, a framework for scalable construction of evidence grounded multimodal rubrics and training data for both a rubric generator and a rubric aware reward model. MM-VeriRubric extracts task aware visual clue profiles, generates four dimensional rubrics over Perception, Grounding, Inference, and Answer, and constructs controlled rejected responses to verify whether each rubric can identify targeted error types. Experiments on multimodal reward modeling benchmarks and downstream Best of N tasks show that MM-VeriRubric improves preference judgment and provides more structured selection signals for multimodal reasoning. Overall, our work moves multimodal rubric reward from textual evaluation criteria toward a visually grounded, verifiable, and trainable data construction framework for future multimodal RLHF and reward model alignment.

Limitations

Reliance on Teacher-Generated Supervision.

MM-VeriRubric relies on strong teacher models to generate visual clue profiles, evidence grounded rubrics, and controlled rejected responses. Although the consistency verification step filters unstable criteria and invalid preference pairs, it cannot fully remove all teacher model errors. If the teacher model misreads the image, overlooks important evidence, or expresses biased evaluation preferences, such errors may be inherited by the generated training data. This limitation suggests that future work can combine multiple teachers, human auditing, or uncertainty based filtering to further improve the

reliability of rubric supervision.

Task and Modality Coverage. This work focuses mainly on image text question answering, visual reasoning, and text response evaluation. The current framework is not yet fully evaluated on video understanding, image generation, image editing, long horizon multimodal interaction, or complex interleaved outputs that mix text with visual content. These settings may require different forms of evidence anchors and different definitions of response quality. Extending MM-VeriRubric to broader multimodal tasks is an important direction for future work.

Verification Boundaries and Downstream Generalization. Controlled rejected responses and consistency filtering allow MM-VeriRubric to test whether a rubric can recognize specified error dimensions, but they do not guarantee that the rubric covers every possible failure mode. A response may contain subtle factual, pragmatic, or safety related issues that are not captured by the generated criteria. In addition, our downstream evaluation focuses on reward model benchmarks and Best of N selection, while full policy optimization with multimodal RLHF remains to be further explored. Future work can study how rubric aware reward models behave when they are used as training rewards for policy models rather than only as inference time judges.

Ethical Considerations

This work aims to improve the interpretability and diagnostic value of multimodal reward modeling, rather than to generate, manipulate, or amplify user facing content. Multimodal data may contain people, sensitive attributes, or privacy related information, so the construction and use of such data should respect the licenses of the original datasets, filter sensitive samples when necessary, and avoid deployment in unaudited high risk scenarios. Rubric aware reward models should not be treated as absolute arbiters without human review, especially in safety critical, medical, or legal settings. We reduce the risk of erroneous reward signals through held out evaluation, controlled rejected responses, consistency filtering, and human inspection, but these procedures cannot fully eliminate misjudgment or inappropriate evaluation criteria.

References

- Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, and 1 others. 2025. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*.
- Guiming Hardy Chen, Shunian Chen, Ruifei Zhang, Junying Chen, Xiangbo Wu, Zhiyi Zhang, Zhihong Chen, Jianquan Li, Xiang Wan, and Benyou Wang. 2024a. Allava: Harnessing gpt4v-synthesized data for lite vision-language models. *arXiv preprint arXiv:2402.11684*.
- Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Conghui He, Jiaqi Wang, Feng Zhao, and Dahua Lin. 2024b. Sharegpt4v: Improving large multi-modal models with better captions. In *European Conference on Computer Vision*, pages 370–387. Springer.
- Yukun Chen, Jiaming Li, Longze Chen, Ze Gong, Jingpeng Li, Zhen Qin, Hengyu Chang, Ancheng Xu, Zhihao Yang, Hamid Alinejad-Rokny, and 1 others. 2026. Rucl: Stratified rubric-based curriculum learning for multimodal large language model reasoning. *arXiv preprint arXiv:2602.21628*.
- Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Naveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, and 1 others. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*.
- Tianrui Guan, Fuxiao Liu, Xiyang Wu, Ruiqi Xian, Zongxia Li, Xiaoyu Liu, Xijun Wang, Lichang Chen, Furong Huang, Yaser Yacoob, and 1 others. 2024. 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*, pages 14375–14385.
- Anisha Gunjal, Anthony Wang, Elaine Lau, Vaskar Nath, Yunzhong He, Bing Liu, and Sean Hendryx. 2025. Rubrics as rewards: Reinforcement learning beyond verifiable domains. *arXiv preprint arXiv:2507.17746*.
- Yushi Hu, Reyhane Askari-Hemmat, Melissa Hall, Emily Dinan, Luke Zettlemoyer, and Marjan Ghazvininejad. 2025. Multimodal rewardbench 2: Evaluating omni reward models for interleaved text and image. *arXiv preprint arXiv:2512.16899*.
- Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, and 1 others. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.
- Dongfu Jiang, Xuan He, Huaye Zeng, Cong Wei, Max Ku, Qian Liu, and Wenhui Chen. 2024. Mantis: Interleaved multi-image instruction tuning. *arXiv preprint arXiv:2405.01483*.

- Zicheng Kong, Dehua Ma, Zhenbo Xu, Alven Yang, Yiwei Ru, Haoran Wang, Zixuan Zhou, Fuqing Bie, Liuyu Xiang, Huijia Wu, and 1 others. 2026. Omnirm: Advancing omni reward modeling via automatic rubric-grounded preference synthesis. *arXiv preprint arXiv:2602.00846*.
- Lei Li, Yuancheng Wei, Zhihui Xie, Xuqing Yang, Yifan Song, Peiyi Wang, Chenxin An, Tianyu Liu, Sujian Li, Bill Yuchen Lin, and 1 others. 2025. Vlrrewardbench: a challenging benchmark for vision-language generative reward models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 24657–24668.
- Dengcan Liu, Fengkai Yang, Xiaohan Wang, Shurui Yan, Jiajun Chai, Jiahao Li, Yikun Ban, Zhendong Mao, Wei Lin, and Guojun Yin. 2026a. Cdrmm: Contrast-driven rubric generation for reliable and interpretable reward modeling. *arXiv preprint arXiv:2603.08035*.
- Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916.
- Rui Liu, Dian Yu, Zhenwen Liang, Yucheng Shi, Tong Zheng, Runpeng Dai, Haitao Mi, Pratap Tokekar, and 1 others. 2026b. Deltarubric: Generative multimodal reward modeling via joint planning and verification. *arXiv preprint arXiv:2605.09269*.
- Tianci Liu, Ran Xu, Tony Yu, Ilgee Hong, Carl Yang, Tuo Zhao, and Haoyu Wang. 2025. Openrubrics: Towards scalable synthetic rubric generation for reward modeling and llm alignment. *arXiv preprint arXiv:2510.07743*.
- Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. 2024. Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. In *International Conference on Learning Representations*, volume 2024, pages 23439–23554.
- Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, and 1 others. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744.
- Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. *Advances in neural information processing systems*, 36:53728–53741.
- Jiacheng Ruan, Wenzhen Yuan, Xian Gao, Ye Guo, Daoxin Zhang, Zhe Xu, Yao Hu, Ting Liu, and Yuzhuo Fu. 2025. Vlrmbench: A comprehensive and challenging benchmark for vision-language reward models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3163–3173.
- Tanik Saikh, Tirthankar Ghosal, Amish Mittal, Asif Ekbal, and Pushpak Bhattacharyya. 2022. Scienceqa: A novel resource for question answering on scholarly articles. *International Journal on Digital Libraries*, 23(3):289–301.
- Wenhao Shi, Zhiqiang Hu, Yi Bin, Junhua Liu, Yang Yang, See Kiong Ng, Lidong Bing, and Roy Ka-Wei Lee. 2024. Math-llava: Bootstrapping mathematical reasoning for multimodal large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 4663–4680.
- Zhiqing Sun, Sheng Shen, Shengcao Cao, Haotian Liu, Chunyuan Li, Yikang Shen, Chuang Gan, Liangyan Gui, Yu-Xiong Wang, Yiming Yang, and 1 others. 2024. Aligning large multimodal models with factually augmented rlhf. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 13088–13110.
- Haoqin Tu, Weitao Feng, Hardy Chen, Hui Liu, Xi-anfeng Tang, and Cihang Xie. 2025. Vilbench: A suite for vision-language process reward modeling. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 6775–6790.
- Vijay Viswanathan, Yanchao Sun, Xiang Kong, Meng Cao, Graham Neubig, and Sherry Wu. 2026. Checklists are better than reward models for aligning language models. *Advances in Neural Information Processing Systems*, 38:114728–114754.
- Fei Wang, Wenxuan Zhou, James Y Huang, Nan Xu, Sheng Zhang, Hoifung Poon, and Muhao Chen. 2024. mdpo: Conditional preference optimization for multimodal large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 8078–8088.
- Weiyun Wang, Zhangwei Gao, Lianjie Chen, Zhe Chen, Jinguo Zhu, Xiangyu Zhao, Yangzhou Liu, Yue Cao, Shenglong Ye, Xizhou Zhu, and 1 others. 2025a. Visualprm: An effective process reward model for multimodal reasoning. *arXiv preprint arXiv:2503.10291*.
- Xiaokun Wang, Peiyu Wang, Jiangbo Pei, Wei Shen, Yi Peng, Yunzhuo Hao, Weijie Qiu, Ai Jian, Tianyidan Xie, Xuchen Song, and 1 others. 2025b. Skywork-vl reward: An effective reward model for multimodal understanding and reasoning. *arXiv preprint arXiv:2505.07263*.
- Tianyi Xiong, Xiyao Wang, Dong Guo, Qinghao Ye, Haoqi Fan, Quanquan Gu, Heng Huang, and Chunyuan Li. 2025. Llava-critic: Learning to evaluate multimodal models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 13618–13628.
- Michihiro Yasunaga, Luke Zettlemoyer, and Marjan Ghazvininejad. 2025. Multimodal rewardbench: Holistic evaluation of reward models for vision language models. *arXiv preprint arXiv:2502.14191*.

Tianyu Yu, Yuan Yao, Haoye Zhang, Taiwen He, Yifeng Han, Ganqu Cui, Jinyi Hu, Zhiyuan Liu, Hai-Tao Zheng, Maosong Sun, and 1 others. 2024. Rlhv-v: Towards trustworthy mllms via behavior alignment from fine-grained correctional human feedback. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13807–13816.

Tianyu Yu, Haoye Zhang, Qiming Li, Qixin Xu, Yuan Yao, Da Chen, Xiaoman Lu, Ganqu Cui, Yunkai Dang, Taiwen He, and 1 others. 2025. Rlaif-v: Open-source ai feedback leads to super gpt-4v trustworthiness. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 19985–19995.

Ya-Qi Yu, Fangyu Hong, Xiangyang Qu, Hao Wang, Gaojie Wu, Qiaoyu Luo, Nuo Xu, Huixin Wang, Wuheng Xu, Yongxin Liao, and 1 others. 2026. Visual preference optimization with rubric rewards. *arXiv preprint arXiv:2604.13029*.

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

Yuhang Zang, Xiaoyi Dong, Pan Zhang, Yuhang Cao, Ziyu Liu, Shengyuan Ding, Shenxi Wu, Yubo Ma, Haodong Duan, Wenwei Zhang, and 1 others. 2025. Internlm-xcomposer2. 5-reward: A simple yet effective multi-modal reward model. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 6547–6563.

Yi-Fan Zhang, Xingyu Lu, Xiao Hu, Chaoyou Fu, Bin Wen, Tianke Zhang, Changyi Liu, Kaiyu Jiang, Kaibing Chen, Kaiyu Tang, and 1 others. 2025a. R1-reward: Training multimodal reward model through stable reinforcement learning. *arXiv preprint arXiv:2505.02835*.

Yi-Fan Zhang, Haihua Yang, Huanyu Zhang, Yang Shi, Zezhou Chen, Haochen Tian, Chaoyou Fu, Hao-tian Wang, Kai Wu, Bo Cui, and 1 others. 2025b. Basereward: A strong baseline for multimodal reward model. *arXiv preprint arXiv:2509.16127*.

Yi-Fan Zhang, Tao Yu, Haochen Tian, Chaoyou Fu, Peiyan Li, Jianshu Zeng, Wulin Xie, Yang Shi, Huanyu Zhang, Junkang Wu, and 1 others. 2025c. Mm-rlhf: The next step forward in multimodal llm alignment. *arXiv preprint arXiv:2502.10391*.

Yongshuo Zong, Ondrej Bohdal, Tingyang Yu, Yongxin Yang, and Timothy Hospedales. 2024. Safety fine-tuning at (almost) no cost: A baseline for vision large language models. *arXiv preprint arXiv:2402.02207*.

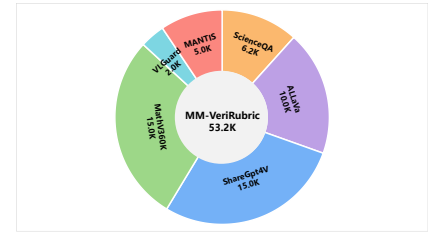


Figure 4: Data sources.

Appendix

A Dataset Construction

A.1 Data Sources and Filtering

We construct the MM-VeriRubric training pool from six image text datasets: ScienceQA, MathV360K, VLGuard, ShareGPT4V, ALLaVA, and MANTIS. These datasets are selected to cover complementary multimodal abilities, including general visual question answering, mathematical visual reasoning, safety related queries, open ended instruction following, and multi image understanding. After preprocessing, each sample is converted into a unified format containing an image or image list, a question, and a reference answer. The data construction pipeline then augments each sample with a task aware visual clue profile, an evidence grounded rubric, controlled rejected responses, and consistency verification results.

To avoid evaluation leakage, we exclude all benchmark samples used for final evaluation from the synthesis pool. In particular, VL-RewardBench, MM-RewardBench, MMRB2-Reasoning, HallusionBench, MathVista, and MMMU-Pro are used only as held out evaluation sets and are not used during rubric synthesis, rejected response construction, or reward model training. We also remove samples with missing images, invalid reference answers, malformed JSON records, or duplicated sample identifiers. The final synthesis pool and data distribution are shown in Fig.4.

A.2 Rubric Synthesis Statistics

Criteria number distribution is shown in Fig.5.

A.3 Consistency Verification Statistics

Consistency Verification Statistics are shown in Fig.6

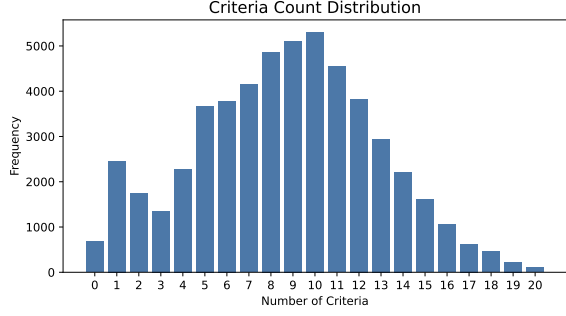


Figure 5: Criteria number distribution.

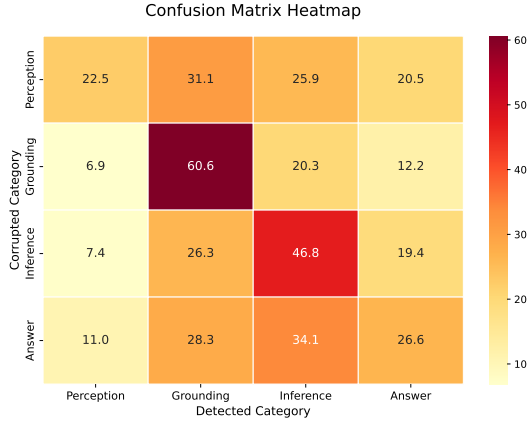


Figure 6: Consistency Verification Statistics

B Implementation Details

B.1 Model Training Details

We train two models in MM-VeriRubric: a rubric generator and a rubric aware reward model. The rubric generator is initialized from an instruction tuned vision language backbone and is trained to produce instance specific evaluation criteria from an image, a question, a reference answer, and a task aware visual clue profile. The rubric aware reward model is initialized from the same backbone family and is trained as a pairwise multimodal judge. It takes an image, a question, a rubric, and two candidate responses as input, then outputs criterion level pass or fail judgments, two response scores, and the final preferred response. In the main experiments, we report two model scales, MM-VeriRubric 4B and MM-VeriRubric 8B, with backbone models set to Qwen3-4B-VL-Instruct and Qwen3-8B-VL-Instruct.

Both models are trained with supervised fine tuning using the multimodal conversation format supported by LLaMA-Factory. For the rubric generator, we use 15,000 training examples converted from the rubric synthesis stage. Each example

contains one user turn with the image, question, reference answer, and generation instruction, and one assistant turn with the target rubric. For the rubric aware reward model, we use 10,000 training examples converted from the consistency verified preference pairs. Each example contains the image, question, rubric, Response A, and Response B as input, and a structured JSON output containing criterion judgments, response scores, the better response label, and a short justification. The order of Response A and Response B is randomly shuffled during data construction to reduce position bias.

Table 4 summarizes the main training configuration.

B.2 Reward Model Evaluation Details

We evaluate MM-VeriRubric as a rubric conditioned pairwise judge on held out reward model benchmarks. For each evaluation sample, the input contains an image I , a question q , two candidate responses y_A and y_B , and a human preference label $g \in \{A, B\}$. In the rubric based evaluation setting, an instance specific rubric $\mathcal{C}$ is first generated for the sample, and the reward model is prompted with $(I, q, \mathcal{C}, y_A, y_B)$. The model is required to output a JSON object containing criterion level judgments, two response scores, and a final field `better_response`. The final pairwise prediction is parsed from this field:

$$\hat{g} = \text{Parse}(f_{\theta}(I, q, \mathcal{C}, y_A, y_B)), \quad \hat{g} \in \{A, B\}.$$

A prediction is counted as correct when $\hat{g} = g$, after accounting for the displayed response order.

The evaluation protocols follow the official settings of each benchmark. For VL-RewardBench, we perform $K = 10$ repeated judgments for each sample and randomly swap the response order in each run. Each run is mapped to one of three statuses: agree, reject, or doesntMatch. The final status of a sample is determined by majority voting over the K runs. Following the official protocol, normalized accuracy is computed over samples whose final status is either agree or reject, while doesntMatch and tie cases are excluded from the denominator:

$$\text{Acc}_{\text{VL}} = \frac{N_{\text{agree}}}{N_{\text{agree}} + N_{\text{reject}}}.$$

For MM-RewardBench, we use $K = 1$ inference by default. In rubric based evaluation, the order of Response A and Response B is randomly

Setting	Rubric Generator	Rubric Aware Reward Model
Backbone	Qwen3-8B-VL-Instruct	Qwen3-8B-VL-Instruct
Training examples	15,000	10,000
Training objective	SFT	SFT
Data format	LLaMA-Factory multimodal conversation	LLaMA-Factory multimodal conversation
Learning rate	1e-5	1e-5
Global batch size	32	32
Epochs	4.0	4.0
Tuning method	full fine tuning	full fine tuning
Hardware	8*H20	8*H20

Table 4: Training configuration for the rubric generator and the rubric aware reward model.

swapped with a fixed seed, and the parsed prediction is mapped back to the original candidate order before comparison with the human label. If the output cannot be parsed into a valid preference label, we follow the benchmark evaluation script and assign a random label with the fixed seed. We report overall pairwise accuracy and macro averaged accuracy over benchmark subcategories:

$$\text{Acc}_{\text{MMRB}} = \frac{1}{N} \sum_{i=1}^N \mathbb{I}[\hat{g}_i = g_i].$$

For MMRB2-Reasoning, we follow the bidirectional evaluation protocol. Each sample is evaluated once in the forward order and once in the reverse order. In the forward run, Response A corresponds to the first candidate and Response B corresponds to the second candidate. In the reverse run, the candidate order is swapped, and the predicted label is flipped back before comparison with the original human label. The final accuracy is computed over all forward and reverse judgments:

$$\text{Acc}_{\text{MMRB2}} = \frac{1}{2N} \sum_{i=1}^N (\mathbb{I}[\hat{g}_i^{\text{fwd}} = g_i] + \mathbb{I}[\text{Flip}(\hat{g}_i^{\text{rev}}) = g_i]).$$

As in MM-RewardBench, invalid JSON outputs or missing preference labels are replaced by a random A or B decision using the fixed evaluation seed.

B.3 Downstream Best of N Details

We evaluate downstream Best of N selection on HallusionBench, MathVista, and MMMU-Pro. These experiments test whether the rubric aware reward model can improve final task performance by selecting a better answer from multiple candidate responses. For each input $x = (I, q)$, we first use a base vision language model to generate one greedy response and N_{max} sampled responses:

$$\mathcal{Y}(x) = \{y_1, \dots, y_{N_{\text{max}}}\}.$$

In our experiments, the base generator is Qwen3-VL-8B-Instruct, and sampled responses are generated with temperature 0.7 and maximum output length 2048 tokens. We evaluate the scaling behavior at

$$N \in \{1, 2, 3, 4, 5\}.$$

The greedy response is evaluated as the baseline, while BoN@ N uses the first N sampled responses as the candidate set.

For each downstream sample, the rubric generator first produces an instance specific rubric $\mathcal{C}_x$. The rubric aware reward model then selects the best response through sequential pairwise elimination. Given a candidate set $\mathcal{Y}_N(x) = \{y_1, \dots, y_N\}$, we randomly shuffle the candidates with a fixed seed and initialize the first candidate as the current best response. The reward model compares the current best response with each remaining candidate under the same rubric:

$$w_t = \text{RM}_\theta(I, q, \mathcal{C}_x, y_{\text{best}}, y_t), \quad w_t \in \{A, B\}.$$

If the model selects B , the challenger replaces the current best response. After $N - 1$ comparisons, the final selected response is

$$\hat{y}_N = \text{Select}_\theta(I, q, \mathcal{C}_x, \mathcal{Y}_N).$$

This sequential procedure requires $N - 1$ reward model calls for each BoN@ N setting and avoids comparing all candidate pairs.

The curve shows the official task metric as N increases from 1 to 5. This analysis is intended to show whether the reward model provides a stable selection signal when the candidate pool becomes larger. A useful reward model should improve over the greedy response and should generally maintain or improve performance as more sampled candidates become available.

The answer extraction and scoring procedure follows the official format of each downstream bench-

mark. For HallusionBench, model outputs are converted into Yes or No labels, and we report aAcc, fAcc, and qAcc. The aAcc metric measures per question accuracy, fAcc requires all questions under the same figure group to be correct, and qAcc requires all variants of the same question group to be correct. For MathVista, answers are normalized according to question type and answer type. Multiple choice answers are mapped to option text, integer and float answers are normalized numerically, and text answers are compared after extraction. For MMMU-Pro, we extract the final option letter from A to J and report overall accuracy, together with subject level and difficulty level breakdowns.

C Additional Experimental Results

C.1 Diagnostic Analysis

We further analyze whether the generated rubrics can localize the intended error dimension. For each retained preference pair that passes consistency verification, we use the injected corrupted dimension as the target dimension d^* . We then compute the score drop of each rubric dimension from the chosen response to the rejected response:

$$\Delta_d = S_d(y^+) - S_d(y^-), \quad S_d(y) = \frac{\text{passed}_d(y)}{\text{total}_d}.$$

The detected failure dimension is defined as the dimension with the largest score drop:

$$\hat{d} = \arg \max_{d \in \mathcal{D}} \Delta_d.$$

This gives a dimension confusion matrix between the intended corrupted dimension d^* and the detected dimension $\hat{d}$. Table 5 reports the row normalized results over 10,605 retained samples.

The matrix shows that Grounding and Inference errors are most clearly localized. When the corruption targets Grounding, the detected dimension is also Grounding in 60.6 percent of retained samples. When the corruption targets Inference, the detected dimension is Inference in 46.8 percent of retained samples. This result suggests that evidence binding errors and reasoning errors produce relatively stable rubric score drops, so the rubric can often identify the intended failure source rather than only ranking the two responses holistically.

The Perception and Answer rows are more diffuse. Perception corruptions are detected as Perception in 22.5 percent of samples, but they are also frequently detected as Grounding or Inference.

This pattern is reasonable because visual perception errors often propagate to downstream evidence binding and reasoning criteria. If a response reads the wrong object, OCR span, number, or visual relation, later claims may also be grounded in the wrong evidence or lead to an invalid inference. Answer corruptions are detected as Answer in 26.6 percent of samples, but they are also often reflected in Grounding and Inference. This indicates that final answer errors are not always isolated at the output layer, since an incorrect final answer may be accompanied by weaker reasoning or evidence use.

Overall, the confusion matrix supports the diagnostic value of MM-VeriRubric. The strong diagonal values for Grounding and Inference show that the rubric has meaningful dimension localization ability. The off diagonal patterns also reveal natural dependencies among error dimensions, especially the propagation from Perception to Grounding and Inference. This analysis suggests that the proposed four dimensional rubric is not merely a scoring template, but also a useful tool for understanding how multimodal errors emerge and interact.

D Prompt Templates

This appendix provides the main prompt templates used in MM-VeriRubric. The templates use placeholder fields such as {query}, {reference_answer}, {visual_clue_profile}, {rubric_json}, and {response_a}. In multimodal calls, the image is provided together with the user message through the vision language model interface.

D.1 Task-Aware Visual Clue Extraction Prompt

Task-Aware Visual Clue Extraction: System Prompt

You are an expert visual analysis assistant.
Your task is to extract a structured Visual Clue Profile from a given image and question pair.

The Visual Clue Profile captures the key visual evidence, reasoning operations, and potential failure modes needed to answer the question correctly.

You must output only a valid JSON object with the following fields:

{

Corrupted / Detected	Perception	Grounding	Inference	Answer
Perception	22.5	31.1	25.9	20.5
Grounding	6.9	60.6	20.3	12.2
Inference	7.4	26.3	46.8	19.4
Answer	11.0	28.3	34.1	26.6

Table 5: Dimension confusion matrix between the injected corrupted dimension and the detected failure dimension. Each row is normalized to 100 percent.

```

"target_entities": [
  "List of key entities in the image that
  are relevant to answering the question."
],
"evidence_anchors": [
  {
    "type": "bbox | ocr_span | object_ref
    | region_desc",
    "value": "Description of the visual
    evidence anchor"
  }
],
"required_operations": [
  "identify | compare | count | rank |
  calculate | read | locate | infer |
  verify"
],
"answer_type": "category_name | number |
short_text | option | yes_no | list",
"query_constraints": [
  "Explicit or implicit constraints from
  the question"
],
"anticipated_failure_modes": [
  "Potential errors a vision language
  model may make"
]
}

Rules:
1. The image is the primary source of truth.
2. Be specific and concrete.
3. For evidence anchors, choose the most
   appropriate type.
4. For anticipated failure modes, focus on
   errors that may affect the final answer.
5. Output only the JSON object.

```

Task-Aware Visual Clue Extraction: User Prompt

Question:
{query}

Please analyze the image and the question above, then extract the Visual Clue Profile as a JSON object.

D.2 Evidence-Grounded Rubric Generation Prompt

Evidence-Grounded Rubric Generation: System Prompt

You are an expert rubric designer for evaluating vision language model responses.

Given an image, a question, a reference answer, and a visual clue profile, generate a set of evaluation criteria. The rubric must be organized into exactly four dimensions.

Dimension definitions:

- 1. Perception**
Tests whether the response correctly reads or identifies raw visual elements from the image. This includes OCR text, numbers, object names, colors, shapes, sizes, counts, attributes, labels, and legends.
- 2. Grounding**
Tests whether the response binds a claim to the correct visual evidence. This includes spatial attribution, entity evidence linkage, and reference resolution.
- 3. Inference**
Tests whether the response performs valid reasoning over correctly perceived and grounded evidence. This includes comparison, calculation, ranking, trend analysis, logical deduction, and multi step reasoning.
- 4. Answer**
Tests whether the final answer satisfies the question requirements in content, format, unit, option selection, completeness, and safety.

Rules for each criterion:

1. Each criterion tests exactly one thing.
2. Each criterion is concise and independently verifiable.
3. Perception, Grounding, and Inference criteria should include a visual evidence anchor.
4. Criteria should avoid redundant coverage across dimensions.

- 5. The image is the primary source of truth.
- 6. Generate more criteria for dimensions where more errors can occur for this specific question.

Output only a valid JSON object:

```
{
  "criteria": [
    {
      "dimension": "Perception | Grounding |
Inference | Answer",
      "criterion": "One atomic, concise, and
verifiable sentence.",
      "anchor": {
        "type": "bbox | ocr_span |
object_ref | region_desc | null",
        "value": "Specific visual evidence,
or null if not applicable"
      },
      "scoring_rule": "binary",
      "failure_mode_tag": "short snake case
tag"
    }
  ]
}
```

Evidence-Grounded Rubric Generation: User Prompt

Question:
{query}

Reference Answer:
{reference_answer}

Visual Clue Profile:
{visual_clue_profile}

Based on the image, the question, the reference answer, and the visual clue profile above, generate the four dimensional evaluation rubric as a JSON object.

D.3 Rubric-Aligned Rejected Response Construction Prompt

Rubric-Aligned Rejected Response Construction: System Prompt

You are an expert at constructing controlled error variants of vision language model responses for rubric quality testing.

Given a correct reference answer to an image based question and an evaluation rubric, construct exactly seven rejected responses with specific and controlled errors.

The four evaluation dimensions are:

- 1. Perception
Corrupt raw visual reading or identification.
Examples include wrong numbers, wrong text, wrong colors, wrong object names, or incorrect counts.
- 2. Grounding
Keep the perceived elements correct but bind them to the wrong region, object, text span, or entity.
- 3. Inference
Keep perception and grounding correct but introduce an invalid calculation, comparison, ranking, or logical step.
- 4. Answer
Keep perception, grounding, and reasoning correct but change only the final stated answer, format, unit, or option.

Required outputs:

Type A. Single dimension errors
Construct four responses. Each response corrupts exactly one dimension: perception, grounding, inference, and answer.

Type B. Mixed dimension errors
Construct two responses. One corrupts perception and inference. The other corrupts grounding and answer.

Type C. Surface rewrite
Rewrite the reference answer with a different surface form but identical meaning.

- Rules:
- 1. Single dimension errors must only corrupt the target dimension.
 - 2. Mixed errors must corrupt exactly the specified two dimensions.
 - 3. The surface rewrite must preserve the same facts, reasoning, and final answer.
 - 4. All responses should sound natural and fluent.
 - 5. Corruptions should be subtle enough to require careful checking.
 - 6. Each response should be approximately the same length as the reference answer.
 - 7. Include a brief corruption description for each rejected response.
 - 8. Output only a valid JSON object.

Output format:

```
{
  "single_dim": {
    "perception": {
      "response": "...",
      "corruption_description": "..."
    },
    "grounding": {
      "response": "...",
      "corruption_description": "..."
    }
  },
  "mixed_dim": {
    "perception_inference": {
      "response": "...",
      "corruption_description": "..."
    },
    "grounding_answer": {
      "response": "...",
      "corruption_description": "..."
    }
  },
  "surface_rewrite": {
    "response": "...",
    "corruption_description": "..."
  }
}
```

```

"inference": {
  "response": "...",
  "corruption_description": "..."
},
"answer": {
  "response": "...",
  "corruption_description": "..."
}
},
"mixed": {
  "perception_inference": {
    "response": "...",
    "corruption_description": "..."
  },
  "grounding_answer": {
    "response": "...",
    "corruption_description": "..."
  }
},
"surface_rewrite": {
  "response": "...",
  "rewrite_description": "..."
}
}

```

Rubric-Aligned Rejected Response Construction: User Prompt

Question:
{query}

Reference Answer:
{reference_answer}

Evaluation Rubric:
{rubric_json}

Please construct the seven rejected responses as a JSON object.

6. Be strict and precise.
7. The labels A and B are arbitrary.
8. After scoring, write a concise justification explaining why the higher scoring response is better. Mention concrete errors using the dimension names Perception, Grounding, Inference, and Answer.

Output only a valid JSON object:

```

{
  "Response A": [1, 0, 1],
  "Response B": [0, 1, 1],
  "justification": "A concise explanation of the preference."
}

```

Consistency Verification: User Prompt

Question:
{query}

Evaluation Criteria:
{criteria_text}

Responses to Score:

[Response A]
{response_a}

[Response B]
{response_b}

Score each response on each criterion. Use 1 for pass and 0 for fail. Output the JSON object.

D.4 Consistency Verification and Reward Model Evaluation Prompts

Consistency Verification: System Prompt

You are an expert evaluator for vision language model responses.

Given an image, a question, an evaluation rubric, and two candidate responses labeled Response A and Response B, score each response on each criterion.

Scoring rules:

1. For each criterion, judge whether the response satisfies it. Use 1 for pass and 0 for fail.
2. Judge each criterion independently.
3. Judge each response independently.
4. Use the image as the primary source of truth when checking Perception and Grounding criteria.
5. Evaluate only what each criterion asks.

Rubric-Aware Reward Model Evaluation: System Prompt

You are a rubric based evaluation judge for image text question answering tasks. Your job is to compare two candidate responses, Response A and Response B, to a question about an image, using a provided set of evaluation criteria.

Input:

1. An image related to the question.
2. A question about the image.
3. A numbered rubric. Each criterion is an atomic and verifiable check item.
4. Two candidate responses.

Evaluation procedure:

For each criterion, judge whether each response satisfies it. Use "pass" if the response clearly satisfies the criterion. Use "fail" if the response does not satisfy or only partially satisfies it.

After evaluating all criteria, count the total number of passed criteria for each response. The response with the higher score is the better one.

Output only a single JSON object:

```
{
  "criteria_judgments": [
    {
      "criterion_index": 1,
      "response_a": "pass or fail",
      "response_b": "pass or fail"
    }
  ],
  "score_a": "<total pass count for Response A>",
  "score_b": "<total pass count for Response B>",
  "better_response": "A or B",
  "justification": "A concise explanation of why the better response is preferred."
}
```

Rules:

1. Evaluate each criterion independently based on the image and the rubric.
2. Do not let length, fluency, or style influence the judgment.
3. If both responses score equally, choose the one that is more precise and concise.
4. If the tie still remains, output A.
5. In the justification, describe concrete errors using dimension names.
6. Output only the JSON object.

Rubric-Aware Reward Model Evaluation: User Prompt

Question:
{query}

Rubric:
{rubric_text}

Response A:
{response_a}

Response B:
{response_b}

Please evaluate both responses against the rubric and output your judgment as JSON.

E Case Study

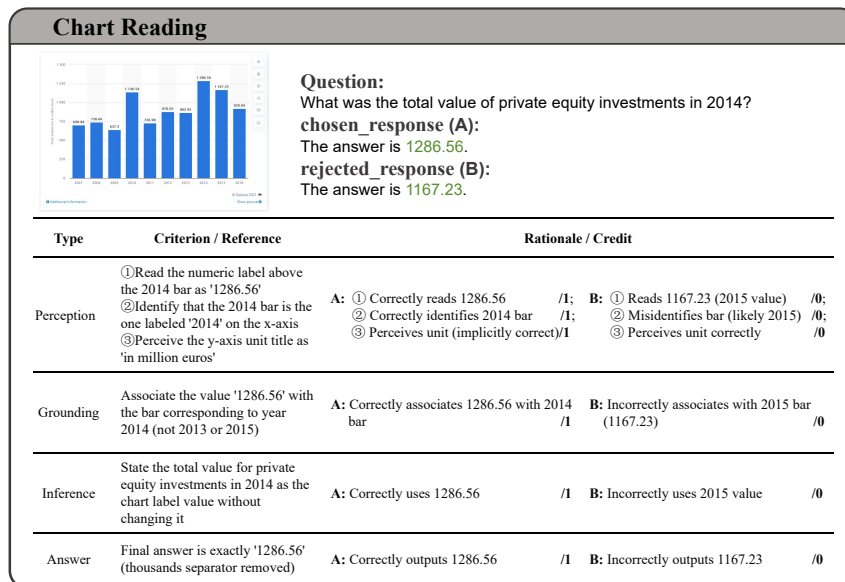


Figure 7: Chart Reading

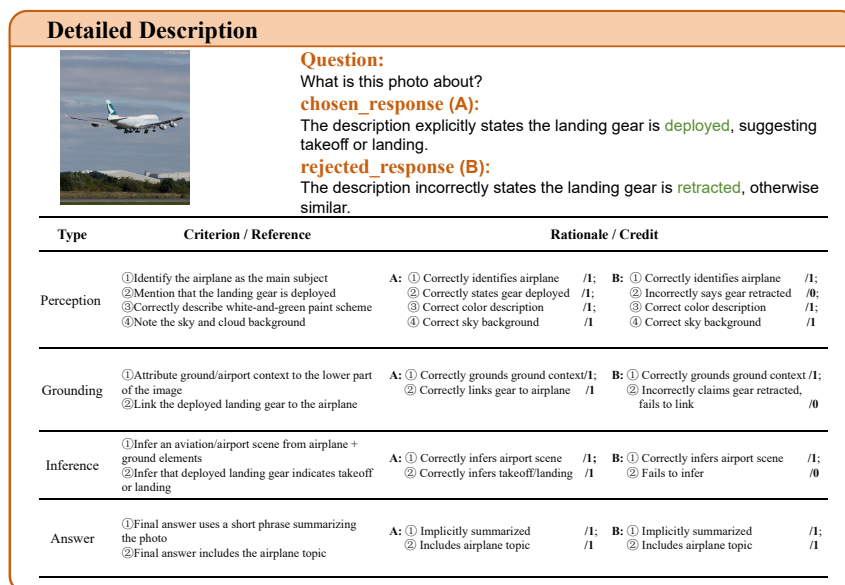


Figure 8: Detailed Description

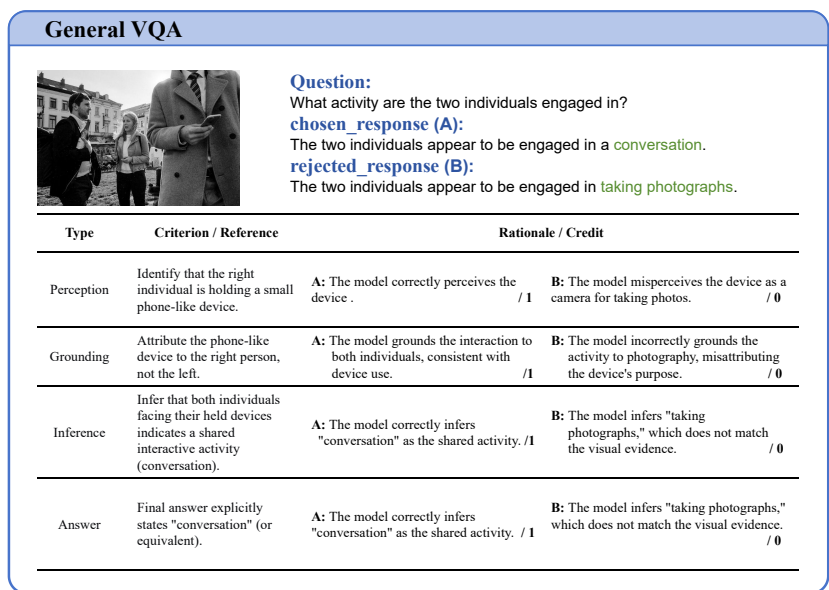


Figure 9: General VQA

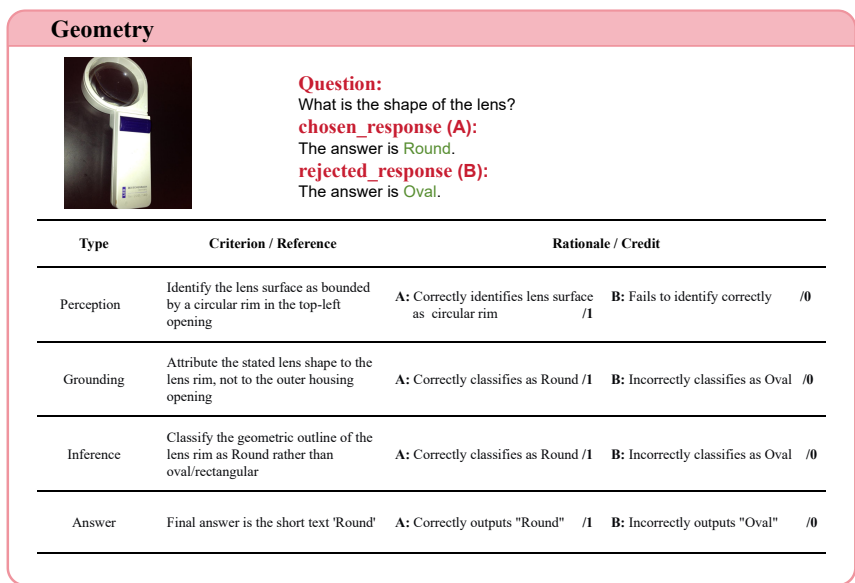


Figure 10: Geometry

Object Presence / Hallucination				
		Question: What elements in this image indicate that the space is designated for a specific gender? chosen_response (A): The image shows a sign with a symbol commonly used to represent men, indicating that the space is designated as a men's restroom. rejected_response (B): The image shows a sign with a symbol commonly used to represent women, indicating that the space is designated as a women's restroom.		
Type	Criterion / Reference	Rationale / Credit		
Perception	① Identify the text label "MEN" on the black restroom sign ② Identify a male/restroom pictogram icon above the word "MEN" ③ Identify the sign as a gender restroom sign	A: ① Correctly identifies "MEN" /1; ② Correctly identifies male pictogram /1; ③ Correctly identifies as gender restroom sign /1	B: ① Fails to identify "MEN" /0; ② Fails to identify male icon /0; ③ Still identifies as gender restroom sign /0	
Grounding	① Link the "MEN" text specifically to the black restroom sign on the right-side wall ② Link the male pictogram icon to the same sign that contains the "MEN" text	A: ① Correctly links "MEN" to the right-side sign /1; ② Correctly links male icon to the same sign /1	B: ① Incorrectly links to female sign /0; ② Incorrectly links /0	
Inference	Infer that the space is designated for men based on the male pictogram and/or the "MEN" label	A: Correctly infers men's space /1	B: Incorrectly infers women's space /0	
Answer	Final answer lists the visible gender-designating elements	A: Correctly lists male-related elements /1	B: Lists incorrect female-related elements /0	

Figure 11: Object Presence or Hallucination