

Visual Preference Optimization with Rubric Rewards

Ya-Qi Yu^{*,†,✉}, Fangyu Hong^{*}, Xiangyang Qu^{*}, Hao Wang^{*},
Gaojie Wu, Qiaoyu Luo, Nuo Xu, Huixin Wang, Wuheng Xu,
Yongxin Liao, Zihao Chen, Haonan Li, Ziming Li, Dezhi Peng,
Minghui Liao, Jihao Wu, Haoyu Ren, Dandan Tu

^{*} Core Contributors [†] Project Leader

Huawei Technologies Co., Ltd.

Project page: <https://github.com/yuyq96/rubrics>

Abstract

The effectiveness of Direct Preference Optimization (DPO) depends on data pairs that capture task-relevant quality differences. Existing data pipelines often rely on *off-policy* perturbations or *coarse* outcome-based signals, which are not well suited to fine-grained visual reasoning. We propose *rDPO*, a rubric-based data pipeline for DPO-based visual preference optimization. For each image-instruction pair, we create a checklist-style rubric of essential and additional criteria to score responses from any possible policies. The instruction-rubric pool is built *offline* and reused during the construction of *on-policy* preference data. On public reward modeling benchmarks, rubric-based prompting massively improves a 30B-A3B judge and brings it close to GPT-5.4. On public downstream benchmarks, rubric-based filtering raises the macro average to 82.69, whereas outcome-based filtering drops it to 75.82 from 81.14. In scalability evaluation on a comprehensive benchmark, rDPO achieves 61.01, markedly outperforming the style-constrained baseline (52.36) and surpassing the 59.48 base model. Together, these results show that visual preference optimization benefits from combining on-policy data construction with instance-specific criterion-level feedback.

1 Introduction

Direct Preference Optimization (DPO) has become a common approach to aligning Vision-Language Models (VLMs), improving quality and reducing hallucinations (Rafailov et al., 2023; Wang et al., 2025a; Yue et al., 2025; Yang et al., 2025a; Team, 2025). Its effectiveness, however, heavily relies on the quality of the underlying preference data. Early pipelines for constructing preference pairs fell into two broad categories: *response-oriented* methods that inject hallucinations or exploit “strong-weak” model outputs (Zhou et al., 2024; Zhao et al., 2025;

Pi et al., 2024; Deng et al., 2024b; Zhu et al., 2024; Deng et al., 2024a; Wang et al., 2024b; Yu et al., 2025b), and *vision-oriented* methods that perturb visual inputs via diffusion noise or image editing (Zhou et al., 2024; Jiang et al., 2025; Wang et al., 2024a; Xie et al., 2024; Luo et al., 2025; Xing et al., 2025; Liu et al., 2025c). While effective in specific scenarios, these methods often yield *off-policy* pairs, causing the constructed data to drift from the target model’s actual response generation patterns on individual inputs.

Recent work has moved toward on-policy and self-improving pipelines. Methods like RLAIFF-V (Yu et al., 2025b) and MMRP (Wang et al., 2024b) group model-generated responses by hallucination count or final correctness, while OPA-DPO (Yang et al., 2025b) and online DPO variants (Liu et al., 2025a; Yu et al., 2025a) repeatedly sample the latest policy during training. Despite reducing data drift, most existing pipelines still rely on coarse-grained reward signals, overlooking crucial nuances in grounding, completeness, and reasoning quality. To enrich supervision, many studies use VLM-as-a-Judge and Reward Models (RMs) to score candidate responses (Sun et al., 2024; Yu et al., 2024a; Lu et al., 2024; Li et al., 2023; Liu et al., 2025d; Zhang et al., 2025b). However, judge quality becomes the primary bottleneck: proprietary models are cost-prohibitive at scale, while open-source judge models (Lee et al., 2024; Wang et al., 2025c; Xiong et al., 2025b; Zhang et al., 2025a; Zang et al., 2025; Wang et al., 2025b), constrained by fixed prompt templates, yield coarse scores that lack the transparency and fine-grained feedback needed to penalize hallucinations or reasoning errors. Rubric-based evaluation enhances LLM alignment by breaking quality into explicit criteria (Hashemi et al., 2024; Starace et al., 2025; Gupta et al., 2025; Viswanathan et al., 2025; Zhou et al., 2025), yet remains underexplored in multi-modal settings. Concurrent work such as Omni-

[✉]E-mail: yuyaqi5@huawei.com

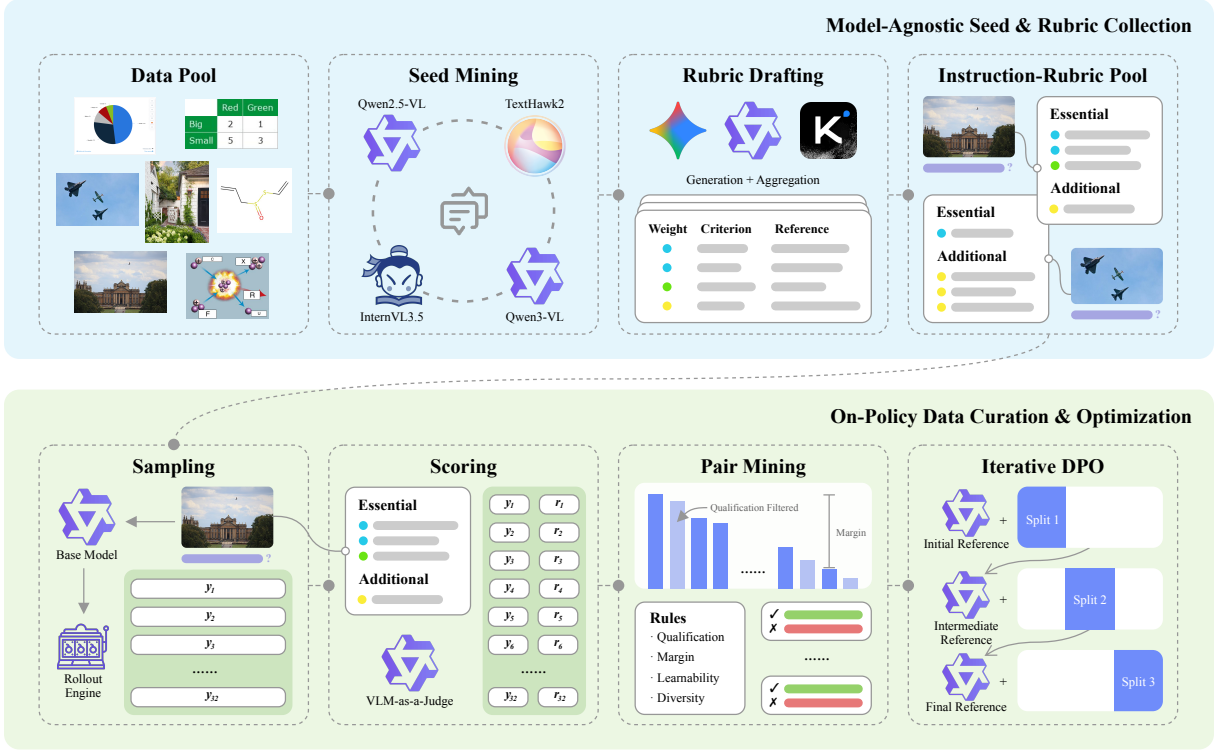


Figure 1: Overview of rDPO. The top lane constructs a model-agnostic instruction-rubric pool by mining challenging seeds via model ensembles and drafting instance-specific rubrics with structured “essential” and “additional” criteria. The bottom lane achieves on-policy data curation and optimization by sampling responses from a target policy, scoring them via rubric-grounded judging, and mining preference pairs.

RRM (Kong et al., 2026) moves in this direction, but it still relies on fixed combinations of predefined evaluation criteria.

We propose rDPO, a rubric-guided data construction pipeline for DPO-based visual preference optimization grounded in *instance-specific* rubrics, as illustrated in Fig. 1. For each image-instruction pair, we build an offline checklist-style rubric that provides structured guidance for evaluating target-policy responses. We then sample on-policy responses, score them with a moderately-sized 30B-A3B open-source judge, and mine preference pairs using criterion-level feedback. Because rubric construction is decoupled from target-policy rollout and scoring, the resulting instruction-rubric pool can be reused across different target models. In this work, we use the constructed preference pairs for iterative DPO. Our evaluation is threefold: judge quality on public reward modeling benchmarks, a method-validation setting where rubric-based filtering is compared with outcome-based filtering under the same on-policy setup, and scaling validation on a comprehensive benchmark.

The main contributions are as follows:

- We introduce a data-centric pipeline that gener-

ates and uses instance-specific multi-criteria rubrics to access on-policy responses, and mines preference data for visual DPO.

- Evaluations on public reward-modeling benchmarks show that instance-specific rubrics improve the judgment quality of a moderately-sized 30B-A3B open-source judge without additional model training.
- We construct 300K reusable instruction-rubric pairs by decoupling offline rubric construction from rollout, scoring, and preference-pair mining for each target policy.
- In method validation, outcome-based filtering reduces the macro average from 81.14 to 75.82, while rDPO raises it to 82.69. In scaling validation, rDPO achieves 61.01, outperforming both the style-constrained baseline (52.36) and the base model (59.48).

2 Related Works

2.1 Multimodal Reward Models

Aligning VLMs with complex human intent has motivated the curation of large-scale multimodal preference datasets. Early efforts relied heavily on human feedback to build datasets such as LLaVA-

RLHF (Sun et al., 2024), RLHF-V (Yu et al., 2024a), MM-RLHF (Zhang et al., 2025b), WildVision (Lu et al., 2024), and VisionArena (Chou et al., 2025). To reduce annotation costs, AI feedback mechanisms followed, as seen in VLFeedback (Li et al., 2023), RLAI-F-V (Yu et al., 2025b), MIA-DPO (Liu et al., 2025d), and LLaVA-Critic (Xiong et al., 2025b). In parallel, specialized visual RMs have been developed as automated evaluators. Early explorations like Prometheus-Vision (Lee et al., 2024), SIMA (Wang et al., 2025c), and LLaVA-Critic (Xiong et al., 2025b) established the foundation. Recent work has further refined RM capabilities: CAREVL (Dai et al., 2025) distills language reward knowledge into visual RMs, SVIP-Reward (Gao et al., 2025) incorporates stepwise visual programs, and comprehensive models like MM-RLHF-Reward (Zhang et al., 2025b), IXC-2.5-Reward (Zang et al., 2025), and Skywork-VL Reward (Wang et al., 2025b) integrate diverse multimodal preference corpora. Despite this progress, existing RMs typically produce scalar scores or rely on rigid prompting templates, and lack the transparency and multi-dimensional granularity needed to guide nuanced visual reasoning.

2.2 Rubric-based Rewards

Rubric-based frameworks are increasingly used in both LLM evaluation and alignment. Task-specific rubrics enable structured assessment of complex, multidimensional tasks (Hashemi et al., 2024; Starace et al., 2025; Arora et al., 2025; Wang et al., 2025d; Akyürek et al., 2025). In post-training, frameworks such as Rubrics-as-Rewards (Gunjal et al., 2025), Rubicon (Huang et al., 2025), and CARMO (Gupta et al., 2025) use rubrics to provide structured reward signals. To further increase rubric discriminability, methods like RLCF (Viswanathan et al., 2025), OpenRubrics (Liu et al., 2025b), and Rubrichub (Li et al., 2026) employ contrastive generation. Dynamic refinement processes are introduced by OnlineRubrics (Rezaei et al., 2025), Auto-Rubric (Xie et al., 2025), and Rubric-ARM (Xu et al., 2026) to mitigate reward over-optimization. RuscaRL (Zhou et al., 2025) further introduces rubrics for both reward modeling and hint-guided rollout generation.

In the multimodal domain, rubrics are under-explored. Recent evaluation benchmarks like JudgeAnything (Pu et al., 2025) and Multi-Crit (Xiong et al., 2025a) have introduced instance-

specific checklists for multi-criteria assessment, highlighting the demand for structured evaluation. For model alignment, the concurrent work Omni-RRM (Kong et al., 2026) represents an early attempt to provide fine-grained feedback. However, it relies on fixed combinations of rubric templates defined in advance.

3 Preliminaries

3.1 DPO

Consider a dataset $\mathcal{D}$ consisting of preference triplets (x, y_c, y_r) , where x denotes the input, y_c and y_r represent the chosen and rejected responses, respectively. Let π_θ represent the policy model parameterized by θ , initialized from a reference model π_0 . The DPO objective is formulated as:

$$\mathcal{L}_{\text{DPO}}(\pi_\theta; \pi_0) = -\log \sigma(\hat{r}_\theta(x, y_c) - \hat{r}_\theta(x, y_r)), \quad (1)$$

where σ is the sigmoid function and the implicit reward is defined as

$$\hat{r}_\theta(x, y) = \beta (\log \pi_\theta(y | x) - \log \pi_0(y | x)). \quad (2)$$

Here, β is the KL penalty coefficient.

The derivation (Rafailov et al., 2023) of the gradient of the DPO loss with respect to θ is:

$$\begin{aligned} \nabla_\theta \mathcal{L}_{\text{DPO}}(\pi_\theta; \pi_0) &= -\beta \underbrace{\sigma(\hat{r}_\theta(x, y_r) - \hat{r}_\theta(x, y_c))}_{\text{scaling by reward margin}} \\ &\quad \cdot \left[\underbrace{\nabla_\theta \log \pi_\theta(y_c | x)}_{\text{increase likelihood of } y_c} - \underbrace{\nabla_\theta \log \pi_\theta(y_r | x)}_{\text{decrease likelihood of } y_r} \right]. \end{aligned} \quad (3)$$

The gradient structure scales the updates based on the reward margin, increasing the likelihood of y_c while penalizing y_r .

3.2 MPO

To integrate preference learning with Supervised Fine-Tuning (SFT), we opt for Mixed Preference Optimization (MPO) (Wang et al., 2024b). This formulation augments the DPO objective with an auxiliary SFT loss:

$$\mathcal{L}_{\text{MPO}}(\pi_\theta; \pi_0) = \mathcal{L}_{\text{DPO}}(\pi_\theta; \pi_0) + \mathcal{L}_{\text{SFT}}(\pi_\theta). \quad (4)$$

The SFT term reinforces the chosen response y_c , acting as a form of rejection sampling:

$$\mathcal{L}_{\text{SFT}}(\pi_\theta) = -\alpha \log \sigma(\log \pi_\theta(y_c | x)), \quad (5)$$

where α is a scaling coefficient. We *exclude the length normalization* to ensure the magnitude consistent with the DPO term.

The resulting gradient:

$$\nabla_{\theta} \mathcal{L}_{\text{SFT}}(\pi_{\theta}) = -\alpha \underbrace{\nabla_{\theta} \log \pi_{\theta}(y_c | x)}_{\text{increase likelihood of } y_c}, \quad (6)$$

increases the weight of the gradient on the chosen response to prevent policy collapse.

3.3 Rubric

To enhance the self-improvement cycle, we sample on-policy responses $y \sim \pi_0(\cdot | x)$ and introduce a rubric-based RM. Scoring against the target model’s own outputs reduces the mismatch introduced by static offline datasets.

Specifically, our framework employs Generative RMs (GenRMs) that evaluate responses against a structured, instance-specific checklist. For each input $x \in \mathcal{X}$, we define a localized rubric encompassing K distinct criteria (e.g., factual accuracy, reasoning steps, and instruction following), denoted as a set:

$$C^x = \{c_1^x, c_2^x, \dots, c_K^x\} \quad (7)$$

Formally, the reward model r_{ϕ} is defined as a function that takes an input $x \in \mathcal{X}$, a generated response $y \in \mathcal{Y}$, and the corresponding rubric $C^x \in \mathcal{C}$ to yield a multi-dimensional feedback vector $\mathbf{s} \in \mathbb{R}^K$. Compared to traditional scalar rewards, this formulation provides greater transparency and interpretability. Furthermore, it enables fine-grained data filtering and facilitates the construction of nuanced preference pairs for subsequent optimization.

4 Methodology

Our pipeline for constructing preference data comprises three stages: (1) *Seed Mining*, which filters high-quality seed instructions from a broader data pool; (2) *Rubric Drafting*, which generates instance-specific rubrics for the selected seeds; and (3) *Preference Data Curation*, which samples on-policy responses and constructs preference pairs for a specific target policy model. The first two stages are *model-agnostic*: the resulting instruction-rubric pool can be reused across different target models. Algorithm 1 summarizes the full data construction pipeline and the subsequent iterative DPO optimization process for the target policy.

Algorithm 1 Rubric-guided preference data construction and iterative DPO.

Require: Source pool $\mathcal{D}_{\text{src}}$ with optional annotations a_x^* ; seed-mining ensemble $\mathcal{E}$; rubric drafters $\mathcal{G}$ and aggregator A ; initial policy π_0 ; rubric-based RM r_{ϕ} ; margin δ ; rounds T

Ensure: Seed pool $\mathcal{S}$, rubric pool $\mathcal{P}$, preference pool $\mathcal{D}$, policy π_T

- 1: $\mathcal{S} \leftarrow \text{Balance}(\{x \in \mathcal{D}_{\text{src}} \mid \text{Disagree}(\mathcal{E}, x)\})$
- 2: $\mathcal{P} \leftarrow \emptyset$
- 3: **for** $x \in \mathcal{S}$ **do**
- 4: $C_m^x \leftarrow G_m(x, a_x^*)$ for each $G_m \in \mathcal{G}$
- 5: $C^x \leftarrow A(x, \{C_m^x\}_m)$; $\mathcal{P} \leftarrow \mathcal{P} \cup \{(x, C^x)\}$
- 6: $\mathcal{D} \leftarrow \emptyset$
- 7: **for** $(x, C^x) \in \mathcal{P}$ **do**
- 8: $\mathcal{Y}_x \leftarrow \text{Sample}(\pi_0(\cdot | x); 32)$
- 9: $\mathcal{Q} \leftarrow \{(y_c, y_r) \in \mathcal{Y}_x^2 \mid \text{Chosen}(y_c) \wedge \text{Rejected}(y_r) \wedge r_{\phi}(x, y_c, C^x) - r_{\phi}(x, y_r, C^x) \geq \delta\}$
- 10: $\mathcal{D} \leftarrow \mathcal{D} \cup \text{TopDiverse}(\mathcal{Q}; 4, 2)$
- 11: Partition $\mathcal{D}$ into $\{\mathcal{D}_1, \dots, \mathcal{D}_T\}$
- 12: **for** $t = 1, \dots, T$ **do**
- 13: $\pi_t \leftarrow \text{DPO}(\pi_{t-1}, \mathcal{D}_t)$
- 14: **return** $\mathcal{S}, \mathcal{P}, \mathcal{D}, \pi_T$

4.1 Seed Mining

To curate a high-quality and challenging set of instructions, we employ a disagreement-based filtering mechanism that operates in two steps:

Large-Scale Rollout. We first conduct a large-scale parallel rollout to collect initial responses. To capture diverse model behaviors during response generation, we use an ensemble of moderately-sized models spanning both dense and Mixture-of-Experts (MoE) architectures: TextHawk2-7B (Yu et al., 2024b), Qwen2.5-VL-7B (Bai et al., 2025), InternVL3.5-30B-A3B (Wang et al., 2025a), and Qwen3-VL-30B-A3B (Team, 2025).

Disagreement-based Filtering. We then apply a reference-free, disagreement-based filtering strategy. We retain only those instances where the ensemble models fail to reach a consensus. Such disagreement often identifies complex, ambiguous, or edge cases that are highly informative for preference optimization. Finally, to maintain domain balance, we apply uniform sampling across all original data sources to construct the final seed set.

4.2 Rubric Drafting

To ensure that the evaluation criteria are mutually exclusive and strictly verifiable by an RM, our construction process adheres to four core principles: (1) *Atomic*: Each criterion targets a single, indivisible key point or sub-query within the original instruction. (2) *Comprehensive*: The criteria list jointly covers all vital dimensions of the user query,

so that no critical aspect of a complete response is overlooked. (3) *Precise*: The evaluation aligns strictly with the user query, avoiding redundant checks or extraneous information. (4) *Objective*: Assessments are grounded in observable facts, empirical evidence, or reference answers, eliminating variance from subjective interpretations.

We explicitly categorize the criteria into two distinct types: *essential* and *additional*. **Essential** criteria capture the core information prioritized by the query; satisfying these is a prerequisite for a conceptually sound response. **Additional** criteria encompass relevant image facts, supplementary knowledge, or intermediate reasoning steps needed to derive a complete and well-supported answer.

Each check item is structured as a triplet comprising a *criterion*, a *reference*, and a fixed-point *weight*. The **criterion** dictates a concrete and observable assertion to be verified. The **reference** serves as the ground truth, strictly derived from image facts. The **weight** quantifies the item’s importance across three discrete levels: 1 (*Auxiliary*) for helpful but non-critical information; 2 (*Important*) for content that significantly enhances the user experience; and 3 (*Key*) for critical elements where any omission or deviation constitutes a clear factual or reasoning error in the response.

To generate rubrics following this schema, we employ frontier reasoning models to generate instance-specific checklists in JSON format (see Appendix A). To account for varying annotation quality across seed datasets, we adapt our generation strategy accordingly:

Expert-Grounded Generation. For clean datasets with trustworthy human annotations, the rubric generation is explicitly conditioned on the provided ground truths. This ensures that the resulting rubrics are firmly anchored in expert knowledge and factual accuracy.

Answer-Agnostic Generation. Conversely, for imperfect datasets characterized by noisy or missing annotations, we deliberately exclude the original answers from the prompt. This prevents the generation process from being misled by reference responses that are inaccurate or incomplete.

Furthermore, relying on a single model for generation risks introducing systematic bias or persistent perception errors. We independently prompt multiple models and synthesize their candidate rubrics into a unified checklist via a separate aggregation prompt that combines the candidate check items.

4.3 Preference Data Curation

We sample multiple responses per instruction from the target policy using the instruction-rubric pool. This rollout yields a diverse candidate set for preference pair construction.

4.3.1 Reward Modeling

Scoring. To evaluate the generated candidates, we employ rubric-grounded VLM-as-a-Judge via zero-shot prompting. For each criterion specified in the rubric, the judge assigns a discrete score $s \in \{0, 0.5, 1\}$, corresponding to *no credit*, *partial credit*, and *full credit*, respectively. The judge is instructed to output the evaluation in JSON format (see Appendix A).

Voting. To reduce the variance and hallucination of VLM-as-a-Judge, we run the scoring process three times independently and adopt the median score for each criterion.

Aggregation. The criterion-level scores are aggregated to compute the overall reward for a given input x and response y . Assume each criterion c_k^x is associated with a reference answer a_k^x and a weight w_k^x , the overall reward is:

$$r_\phi(x, y, C^x) = \sum_k w_k^x \cdot s_\phi(x, y, c_k^x, a_k^x), \quad (8)$$

where ϕ denotes the parameters of the judge.

4.3.2 Pair Mining

Once all candidate responses are scored, we construct the final preference pairs by applying the following filtering criteria. This guarantees that the reward margin reflects a genuine and interpretable difference in quality. The pairing process strictly adheres to the following rules:

- **Rule 1 (Chosen Qualification):** The chosen response y_c must receive at least partial credit on every essential criterion, and at most one essential criterion may receive partial credit. Equivalently, no essential criterion may receive a score of 0, at most one may receive 0.5, and all remaining essential criteria must receive 1. Furthermore, responses exhibiting repetition loops or language mixing are discarded.
- **Rule 2 (Rejected Qualification):** The rejected response y_r must exhibit definitive flaws. It must either fail ($s = 0$) at least one essential criterion or receive partial credit ($s = 0.5$) on at least two essential criteria.
- **Rule 3 (Margin Constraint):** To ensure a meaningful qualitative gap, the overall reward

difference must satisfy a margin constraint: $r_\phi(x, y_c, C^x) - r_\phi(x, y_r, C^x) \geq \delta$, where $\delta > 0$ is an empirical margin hyperparameter to ensure a distinct quality gap.

- **Rule 4 (Maximum Learning Capacity):** Among all valid pairs satisfying the above criteria for a given instruction, we select the top four pairs (y_c, y_r) with the largest overall reward margins between the chosen and rejected responses for that instruction.
- **Rule 5 (Diversity Control):** To prevent the policy model from overfitting, each unique response is allowed to appear a maximum of twice across all preference pairs included in the final training dataset.

When to Use DPO Rather than GRPO? In this work, we choose DPO for three practical reasons. First, it improves efficiency: rollouts and rubric judging can be highly parallelized, avoiding synchronization bubbles and improving inference utilization. Second, it simplifies auditing: reviewers need only verify that the chosen response is clearly better than the rejected one. Third, it provides greater control: quality and style preferences remain explicit and fixed throughout training, reducing exposure to reward hacking and unintended length or style drift. Nevertheless, Group Relative Policy Optimization (GRPO) (Shao et al., 2024) can be attractive for directly incorporating fine-grained rubric rewards into policy optimization, provided that reward hacking and unintended preference drift can be adequately controlled.

5 Experiment

5.1 Results on Reward Modeling

Benchmarks. To validate the effectiveness of the proposed rubrics in multimodal reward modeling, we compare it against visual judges across several preference benchmarks: Multimodal RewardBench (MM-RB) (Yasunaga et al., 2025), VL-RewardBench (VL-RB) (Li et al., 2025), MLLM-as-a-Judge (MaaJ) (Chen et al., 2024a), VisionArena-Battle (VA-B) (Chou et al., 2025), and WildVision-Battle (WV-B) (Lu et al., 2024). Across all benchmarks, we exclude tie samples to focus on clear preferences. Furthermore, we sub-sample 1,200 pairs from VA-B and WV-B, respectively, to optimize evaluation efficiency.

Baselines. To ensure a rigorous evaluation, we benchmark our approach against a wide range of state-of-the-art models from three distinct cate-

gories: (1) Proprietary Frontier VLMs, such as Claude 4.6 Opus (Anthropic, 2026), GPT-5.4 (OpenAI, 2026), and the Gemini series (Gemini Team, 2025); (2) Open-source Generalist VLMs, specifically the Qwen3-VL series (Team, 2025); and (3) Open-source Specialist RMs, including IXC-2.5-Reward (Zang et al., 2025) and Skywork-VL Reward (Wang et al., 2025b).

Potential Data Contamination. We mark results as potentially contaminated when they show abnormal benchmark-specific jumps inconsistent with the model’s performance on other reward-modeling benchmarks, together with unusually balanced recall across answer positions. Because the affected benchmarks are widely used human-preference sources and may overlap with training data, we exclude macro averages containing these entries while retaining the unaffected subset averages.

Table 1 shows that rubric prompting improves both Qwen3-VL judges over their vanilla prompts, whereas CoT prompting does not consistently help. For Qwen3-VL-30B-A3B, rubric raises the overall macro average from 71.73 to 74.83. For Qwen3-VL-32B, it raises the macro average from 73.90 to 75.05. These results validate that instance-specific rubrics consistently enhance zero-shot judge quality, lifting Qwen3-VL judges and bringing the A3B model close to GPT-5.4 on the evaluation suite, with macro averages of 74.83 and 74.91, respectively, across the same benchmarks.

However, CoT prompting exhibits mixed effects. It raises the macro average for the 32B judge from 73.90 to 74.45, but lowers that of the 30B-A3B judge from 71.73 to 69.18. Our case analysis suggests that free-form reasoning can overvalue longer or more elaborate explanations despite factual errors, particularly for the smaller 30B-A3B judge. Rubric prompting mitigates this failure mode by constraining evaluation to an instance-specific checklist.

An unexpected finding is the MaaJ decline, from 70.74 to 69.13 for the 30B-A3B judge and from 72.18 to 68.93 for the 32B judge. Our case analysis finds pairs where neither response is correct, but a rubric that does not explicitly prefer abstention can favor a relevant yet incorrect answer over an unusable response.

5.2 Results on Preference Optimization

Data Curation. We construct two downstream preference datasets and study them in two complementary settings. The first is a method-validation

Table 1: Comprehensive evaluation results across diverse preference benchmarks. Accuracy (%) is reported. + *Rubric* denotes our proposed strategy. *Macro* presents a triplet: (Full Average / Average excluding MM-RB & VL-RB / Average excluding VA-B & WV-B). Gray numbers indicate potential data contamination. If a model has contaminated entries, the corresponding macro averages are ignored (-). The best performance in each section is **bolded**, and the second best is underlined.

Model Evaluator	MM-RB	VL-RB	MaaJ	VA-B	WV-B	Macro (Full / - RB / - B)
<i>Proprietary Frontier VLMs</i>						
Claude 4.6 Opus	82.23	<u>74.42</u>	71.39	78.83	74.00	76.17 / 74.74 / 76.01
GPT-5.4	<u>81.91</u>	78.27	<u>70.35</u>	74.25	69.75	<u>74.91</u> / 71.45 / 76.84
Gemini 3.1 Pro	88.79	86.45	66.56	76.25	73.33	- / 72.05 / -
Gemini 3.0 Flash	87.90	85.89	66.60	<u>76.66</u>	<u>73.42</u>	- / <u>72.23</u> / -
Gemini 3.1 Flash-Lite	82.25	81.40	64.31	74.33	67.00	- / 68.55 / -
<i>Open-source Generalist VLMs</i>						
Qwen3-VL-30B-A3B-Instruct	75.82	65.52	70.74	74.08	72.50	71.73 / 72.44 / 70.69
+ <i>CoT</i>	72.49	60.22	70.20	72.17	70.83	69.18 / 71.07 / 67.64
+ <i>Rubric</i>	<u>82.06</u>	73.54	69.13	<u>75.50</u>	<u>73.92</u>	<u>74.83</u> / 72.85 / <u>74.91</u>
Qwen3-VL-32B-Instruct	79.35	70.73	72.18	75.42	71.83	73.90 / 73.14 / 74.09
+ <i>CoT</i>	81.19	72.15	<u>71.83</u>	74.58	72.50	74.45 / 72.97 / 75.06
+ <i>Rubric</i>	83.02	<u>72.49</u>	68.93	76.83	74.00	75.05 / 73.25 / 74.81
<i>Open-source Specialist RMs</i>						
IXC-2.5-Reward	<u>69.12</u>	<u>66.64</u>	69.15	86.25	89.83	- / - / <u>68.30</u>
Skywork-VL Reward	74.25	73.54	<u>59.93</u>	71.67	65.08	68.89 / 65.56 / 69.24

setting designed to isolate the effect of rubric-based data curation. Here we use the training splits of AI2D (Kembhavi et al., 2016), ChartQA (Masry et al., 2022), and M3CoT (Chen et al., 2024b) as seed sources, evaluate the resulting policies on the corresponding test sets, and compare outcome-based filtering against rubric-based filtering under the same base policy and training recipe. The second is a scaling-validation setting designed to test the pipeline in a broader multi-task regime. Here we use the seed pool described in Section 4.1 and evaluate the resulting policy on our comprehensive in-house benchmark under response style constraints. For each seed, we independently generate one rubric draft with Gemini 3.0 Flash and one with Qwen3-VL-235B, and use Gemini 3.0 Flash to aggregate the two drafts. We sample 32 candidate responses per prompt from the target policy, setting the temperature to 1.0. We then filter the responses with the procedure in Section 4.3.2 to construct preference pairs. Prompts that yield no valid pairs are discarded.

Training Details. We select Qwen3-VL-30B-A3B-Instruct as the initial policy. The training is driven by Adam optimizer with a constant learning rate of 1×10^{-6} , $\beta = 0.5$, and $\alpha = 0$. We use a global

Table 2: Results for the method-validation setting. Δ denotes absolute changes from the base model.

Policy	AI2D	ChartQA	M3CoT	Macro
Base model	84.10	82.32	77.00	81.14
+ <i>Outcome-based</i>	78.53	75.48	73.45	75.82
Δ	-5.57	-6.84	-3.55	-5.32
+ <i>Rubric-based</i>	85.95	83.02	79.11	82.69
Δ	+1.85	+0.70	+2.11	+1.55

batch size of 128 for the method-validation setting and 512 for the scaling-validation setting. For the latter, we further divide the constructed dataset into three splits and apply iterative DPO.

Table 2 summarizes the method-validation setting. Outcome-based filtering lowers performance on all three tasks relative to the base model: AI2D drops from 84.10 to 78.53, ChartQA from 82.32 to 75.48, and M3CoT from 77.00 to 73.45, reducing the macro average from 81.14 to 75.82. In contrast, rubric-based filtering improves all three metrics over the base model, reaching 85.95 on AI2D, 83.02 on ChartQA, and 79.11 on M3CoT, with a macro average of 82.69. These results suggest that on-policy data alone is not sufficient in this setting, and the gains depend on pairing those rollouts with instance-specific criterion-level feedback.

Table 3: Results for the scaling-validation setting on the comprehensive benchmark. *Prompting* and *rDPO* denotes the baseline and our model evaluated under the response style constraints. Δ represents absolute gains over the *Prompting* baseline.

Policy	Basic		Advanced		Avg.
	Perc.	Under.	Know.	Reas.	
Base model	56.26	74.10	71.94	51.53	59.48
+ Prompting	55.67	68.57	61.02	42.86	52.36
+ rDPO	62.82	77.00	71.27	51.62	61.01
Δ	+7.15	+8.43	+10.25	+8.76	+8.65

As shown in Table 3, the default baseline achieves an average score of 59.48, though it tends to produce overly verbose responses with excessive Markdown formatting. Attempting to control this verbosity via concise system prompts causes the score to drop to 52.36. In contrast, by training on conciseness-constrained on-policy data, rDPO reaches a superior score of 61.01 and delivers an 8.65 point improvement over the baseline that uses concise system prompts.

Compared with the original base model, rDPO is slightly higher on average (61.01 vs. 59.48) and improves perception and understanding, while the knowledge dimension remains slightly lower than the base model (71.27 vs. 71.94). Together with the method-validation results, this supports the view that on-policy data becomes more effective when paired with rubric-guided criterion-level scoring. In the scaling-validation setting, rDPO recovers the drop introduced by prompt-only conciseness enforcement and slightly improves average score over the base model.

5.3 Ablation Study

Ablation Study of Policy Alignment. To investigate the necessity of on-policy data during alignment, we analyze preference optimization under severe distribution shifts using the MMRP dataset (Wang et al., 2024b), which was originally gathered for the InternVL series (Wang et al., 2025a). Using Qwen3-VL-30B-A3B-Instruct as the base model, we evaluate the following three training configurations:

- **DPO (Off-Policy):** Standard DPO on MMRP with the base checkpoint as π_0 and KL penalty coefficient $\beta = 0.001$, leaving the full distributional mismatch unresolved.
- **MPO (Off-Policy):** MPO training uses an SFT scaling coefficient $\alpha = 0.001$, where the SFT

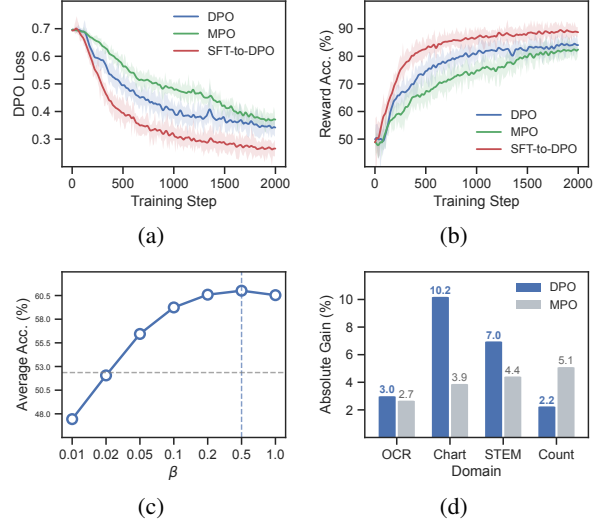


Figure 2: Training dynamics and hyperparameter ablation. (a) DPO loss and (b) reward accuracy for three configurations on off-policy data. (c) Effect of varying the KL penalty coefficient β . (d) Effect of the SFT loss, broken down by task domain.

term supervises chosen responses.

- **SFT-to-DPO (On-Policy Approx.):** A two-stage progressive approach designed to simulate on-policy optimization. The model first undergoes SFT on the chosen responses. The resulting checkpoint then serves as the reference model for subsequent DPO training, thereby narrowing the distribution gap between the policy and the preference data.

Figs. 2a and 2b reveal distinct convergence behaviors. Standard DPO converges slowly due to gradient instability caused by mismatched reference policy and off-policy data distributions. MPO exhibits non-uniform loss decay—fast, then slow, then fast again—reflecting continuous policy shifts. Conversely, SFT-to-DPO is the most stable, achieving the lowest final loss and the highest reward accuracy among the three training configurations.

Our analysis reveals that DPO is inherently sensitive to distribution shift, which degrades the accuracy of the implicit reward. The “inflection point” observed during MPO training likely indicates an initial phase where the model first resolve formatting discrepancies before it can effectively learn from preference signals. The SFT-to-DPO pipeline avoids this bottleneck by ensuring the reference policy matches the data distribution, leading to more stable gradient updates throughout training.

Ablation on the KL Penalty Coefficient. Contrary to previous literature, we empirically find that

Table 4: Results for ablation studies. *rDPO* denotes the full pipeline (baseline for the ablation). Subsequent rows denote variants where specific alignment modules are either removed or altered.

Settings	Basic		Advanced		Avg.
	Perc.	Under.	Know.	Reas.	
rDPO	62.82	77.00	71.27	51.62	61.01
- Iterative	63.02	75.73	71.27	48.62	59.27
- Qualification	60.76	75.20	67.93	48.62	58.36
Rand. margin	63.22	77.69	71.27	50.74	60.76
Min. margin	62.03	76.22	69.71	50.52	59.96
- Diversity	61.95	76.73	70.14	49.00	59.34

the KL penalty coefficient reaches an optimal trade-off at $\beta = 0.5$ (Fig. 2c). Maintaining a substantial penalty anchors the policy to the reference distribution, mitigating policy collapse by preventing the reduction of chosen-response likelihoods (Pal et al., 2024) or overfitting to the implicit reward during preference optimization (Park et al., 2024).

Ablation on the SFT Scaling Coefficient. Figure 2d compares the greatest performance gain of DPO and MPO across different domains. We observe that the effect of the SFT loss depends on the specific task. In most cases, MPO leads to worse results. Although SFT loss improves training stability, it may cause overfitting on low-complexity samples. However, MPO remains effective for tasks where the base model has undesirable priors. A typical example is counting, where giving the final answer before counting step-by-step reduces accuracy. In these cases, the SFT loss acts as a constraint that aligns the model with the high-quality distribution obtained through rejection sampling.

Ablation on Iterative Optimization. We compare the standard single-stage DPO with an iterative paradigm (i.e., the full rDPO pipeline), in which the preference dataset is partitioned into three sequential training splits. As shown in Table 4, iterative DPO achieves better overall performance (61.01 vs. 59.27). In this setup, progressively updating the reference model on partitioned data helps calibrate the implicit reward, leading to improved overall performance on the comprehensive benchmark.

Ablation on Essential Qualification. We compare our approach against a baseline lacking the essential/additional criteria distinction (Rules 1 & 2). As shown in Table 4, enforcing this qualification rule improves average performance from 58.36 to 61.01. This pattern suggests that separating funda-

mental errors from additional criteria helps produce cleaner preference pairs in our setup.

Ablation on Learning Capacity. We compare models trained on datasets constructed using maximum, minimum, and random reward margins (Rule 3 & 4, all strictly bounded by $\delta = 5$ to prevent tie samples). Table 4 shows that selecting pairs with the maximum reward margin gives the best overall performance (61.01 vs. 59.96 vs. 60.76). In this setting, larger reward margins appear to provide a more useful training signal than smaller or randomly selected margins.

Ablation on Diversity Control. We evaluate the impact of limiting response frequency (Rule 5). Removing this constraint degrades the average performance to 59.34, suggesting that reducing exposure to repetitive response patterns enhances downstream performance.

6 Conclusion

We presented rDPO, a data-centric pipeline for DPO-based visual preference optimization using instance-specific rubrics. The pipeline separates offline instruction-rubric pool construction from rollout, scoring, and on-policy data curation. On public RM benchmarks, rubric prompting improves both Qwen3-VL judges over their vanilla prompts, whereas CoT prompting does not consistently help. It brings the moderately-sized 30B-A3B judge close to GPT-5.4 on the reported suite (74.83 vs. 74.91). In the method-validation setting, under the same on-policy rollout setup, outcome-based filtering lowers the macro average from 81.14 to 75.82, while rubric-based filtering raises it to 82.69. When evaluating scalability on our comprehensive benchmark, rDPO achieves 61.01, markedly outperforming the style-constrained baseline (52.36) and surpassing the 59.48 base model. Taken together, these results suggest that effective visual preference optimization requires not only on-policy data, but also instance-specific criterion-level feedback.

7 Limitations

The current pipeline converts criterion-level rubric scores into preference pairs. DPO then optimizes only these pairs, leaving the graded criterion scores outside its objective. This conversion may discard fine-grained information about criterion satisfaction. A natural extension is to investigate online GRPO with rubric rewards and mechanisms that control reward hacking and preference drift.

References

- Afra Feyza Akyürek, Advait Gosai, Chen Bo Calvin Zhang, Vipul Gupta, Jaehwan Jeong, Anisha Gunjal, Tahseen Rabbani, Maria Mazzone, David Randolph, Mohammad Mahmoudi Meymand, Gurshaan Chattha, Paula Rodriguez, Diego Mares, Pavit Singh, Michael Liu, Subodh Chawla, Pete Cline, Lucy Ogaz, Ernesto Hernandez, and 5 others. 2025. [Prbench: Large-scale expert rubrics for evaluating high-stakes professional reasoning](#). *CoRR*, abs/2511.11562.
- Anthropic. 2026. Introducing claude opus 4.6. <https://www.anthropic.com/news/introducing-claude-opus-4-6>. Released February 5, 2026.
- Rahul K. Arora, Jason Wei, Rebecca Soskin Hicks, Preston Bowman, Joaquin Quiñero Candela, Foivos Tsimplouras, Michael Sharman, Meghan Shah, Andrea Vallone, Alex Beutel, Johannes Heidecke, and Karan Singhal. 2025. [Healthbench: Evaluating large language models towards improved human health](#). *CoRR*, abs/2505.08775.
- Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibao Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Ming-Hsuan Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, and 8 others. 2025. [Qwen2.5-vl technical report](#). *CoRR*, abs/2502.13923.
- Dongping Chen, Ruoxi Chen, Shilin Zhang, Yaochen Wang, Yinuo Liu, Huichi Zhou, Qihui Zhang, Yao Wan, Pan Zhou, and Lichao Sun. 2024a. [Mllm-as-a-judge: Assessing multimodal llm-as-a-judge with vision-language benchmark](#). In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*, volume 235 of *Proceedings of Machine Learning Research*, pages 6562–6595. PMLR / OpenReview.net.
- Qiguang Chen, Libo Qin, Jin Zhang, Zhi Chen, Xiao Xu, and Wanxiang Che. 2024b. [M³cot: A novel benchmark for multi-domain multi-step multi-modal chain-of-thought](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024*, pages 8199–8221. Association for Computational Linguistics.
- Christopher Chou, Lisa Dunlap, Koki Mashita, Krishna Mandal, Trevor Darrell, Ion Stoica, Joseph E. Gonzalez, and Wei-Lin Chiang. 2025. [Visionarena: 230k real world user-vlm conversations with preference labels](#). In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2025, Nashville, TN, USA, June 11-15, 2025*, pages 3877–3887. Computer Vision Foundation / IEEE.
- Muzhi Dai, Jiashuo Sun, Zhiyuan Zhao, Shixuan Liu, Rui Li, Junyu Gao, and Xuelong Li. 2025. [From captions to rewards \(carevl\): Leveraging large language model experts for enhanced reward modeling in large vision-language models](#). In *Proceedings of the 33rd ACM International Conference on Multimedia, MM 2025, Dublin, Ireland, October 27-31, 2025*, pages 4972–4981. ACM.
- Shijian Deng, Wentian Zhao, Yu-Jhe Li, Kun Wan, Daniel Miranda, Ajinkya Kale, and Yapeng Tian. 2024a. [Efficient self-improvement in multimodal large language models: A model-level judge-free approach](#). *CoRR*, abs/2411.17760.
- Yihe Deng, Pan Lu, Fan Yin, Ziniu Hu, Sheng Shen, Quanquan Gu, James Y. Zou, Kai-Wei Chang, and Wei Wang. 2024b. [Enhancing large vision language models with self-training on image comprehension](#). In *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*.
- Minghe Gao, Xuqi Liu, Zhongqi Yue, Yang Wu, Shuang Chen, Juncheng Li, Siliang Tang, Fei Wu, Tat-Seng Chua, and Yueting Zhuang. 2025. [Benchmarking multimodal cot reward model stepwise by visual program](#). *CoRR*, abs/2504.06606.
- Google Gemini Team. 2025. [Gemini 3: A family of highly capable multimodal reasoning models](#). *arXiv preprint arXiv:2512.03267*.
- Anisha Gunjal, Anthony Wang, Elaine Lau, Vaskar Nath, Bing Liu, and Sean Hendryx. 2025. [Rubrics as rewards: Reinforcement learning beyond verifiable domains](#). *CoRR*, abs/2507.17746.
- Taneesh Gupta, Shivam Shandilya, Xuchao Zhang, Rahul Madhavan, Supriyo Ghosh, Chetan Bansal, Huaxiu Yao, and Saravan Rajmohan. 2025. [CARMO: dynamic criteria generation for context aware reward modelling](#). In *Findings of the Association for Computational Linguistics, ACL 2025, Vienna, Austria, July 27 - August 1, 2025*, volume ACL 2025 of *Findings of ACL*, pages 2202–2261. Association for Computational Linguistics.
- Helia Hashemi, Jason Eisner, Corby Rosset, Benjamin Van Durme, and Chris Kedzie. 2024. [Llm-rubric: A multidimensional, calibrated approach to automated evaluation of natural language texts](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024*, pages 13806–13834. Association for Computational Linguistics.
- Zenan Huang, Yihong Zhuang, Guoshan Lu, Zeyu Qin, Haokai Xu, Tianyu Zhao, Ru Peng, Jiaqi Hu, Zhanming Shen, Xiaomeng Hu, Xijun Gu, Peiyi Tu, Jiabin Liu, Wenyu Chen, Yuzhuo Fu, Zhiting Fan, Yanmei Gu, Yuanyuan Wang, Zhengkai Yang, and 2 others. 2025. [Reinforcement learning with rubric anchors](#). *CoRR*, abs/2508.12790.
- Songtao Jiang, Yan Zhang, Ruizhe Chen, Tianxiang Hu, Yeying Jin, Qinglin He, Yang Feng, Jian Wu, and Zuozhu Liu. 2025. [Modality-fair preference optimization for trustworthy MLLM alignment](#). In

- Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence, IJCAI 2025, Montreal, Canada, August 16-22, 2025*, pages 403–411. ijcai.org.
- Aniruddha Kembhavi, Mike Salvato, Eric Kolve, Min Joon Seo, Hannaneh Hajishirzi, and Ali Farhadi. 2016. [A diagram is worth a dozen images](#). In *Computer Vision - ECCV 2016 - 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part IV*, Lecture Notes in Computer Science, pages 235–251. Springer.
- Zicheng Kong, Dehua Ma, Zhenbo Xu, Alven Yang, Yiwei Ru, Haoran Wang, Zixuan Zhou, Fuqing Bie, Liyu Xiang, Huijia Wu, Jian Zhao, and Zhaofeng He. 2026. [Omni-rrm: Advancing omni reward modeling via automatic rubric-grounded preference synthesis](#). *CoRR*, abs/2602.00846.
- Seongyun Lee, Seungone Kim, Sue Hyun Park, Geewook Kim, and Minjoon Seo. 2024. [Prometheus-vision: Vision-language model as a judge for fine-grained evaluation](#). In *Findings of the Association for Computational Linguistics, ACL 2024, Bangkok, Thailand and virtual meeting, August 11-16, 2024*, volume ACL 2024 of *Findings of ACL*, pages 11286–11315. Association for Computational Linguistics.
- 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. 2025. [Vl-rewardbench: A challenging benchmark for vision-language generative reward models](#). In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2025, Nashville, TN, USA, June 11-15, 2025*, pages 24657–24668. Computer Vision Foundation / IEEE.
- Lei Li, Zhihui Xie, Mukai Li, Shunian Chen, Peiyi Wang, Liang Chen, Yazheng Yang, Benyou Wang, and Lingpeng Kong. 2023. [Silkie: Preference distillation for large visual language models](#). *CoRR*, abs/2312.10665.
- Sunzhu Li, Jiale Zhao, Miteto Wei, Huimin Ren, Yang Zhou, Jingwen Yang, Shunyu Liu, Kaike Zhang, and Wei Chen. 2026. [Rubrichub: A comprehensive and highly discriminative rubric dataset via automated coarse-to-fine generation](#). *CoRR*, abs/2601.08430.
- Shujun Liu, Siyuan Wang, Zejun Li, Jianxiang Wang, Cheng Zeng, and Zhongyu Wei. 2025a. [Ovip: On-line vision-language preference learning](#). *CoRR*, abs/2505.15963.
- Tianci Liu, Ran Xu, Tony Yu, Ilgee Hong, Carl Yang, Tuo Zhao, and Haoyu Wang. 2025b. [Open-rubrics: Towards scalable synthetic rubric generation for reward modeling and LLM alignment](#). *CoRR*, abs/2510.07743.
- Wenqi Liu, Xuemeng Song, Jiaxi Li, Yinwei Wei, Na Zheng, Jianhua Yin, and Liqiang Nie. 2025c. [Mitigating hallucination through theory-consistent symmetric multimodal preference optimization](#). *CoRR*, abs/2506.11712.
- Ziyu Liu, Yuhang Zang, Xiaoyi Dong, Pan Zhang, Yuhang Cao, Haodong Duan, Conghui He, Yuanjun Xiong, Dahua Lin, and Jiaqi Wang. 2025d. [MIA-DPO: multi-image augmented direct preference optimization for large vision-language models](#). In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net.
- Yujie Lu, Dongfu Jiang, Wenhui Chen, William Yang Wang, Yejin Choi, and Bill Yuchen Lin. 2024. [Wild-vision: Evaluating vision-language models in the wild with human preferences](#). In *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*.
- Tianghe Luo, Ang Cao, Gunhee Lee, Justin Johnson, and Honglak Lee. 2025. [Probing visual language priors in vlms](#). In *Forty-second International Conference on Machine Learning, ICML 2025, Vancouver, BC, Canada, July 13-19, 2025*, volume 267 of *Proceedings of Machine Learning Research*. PMLR / OpenReview.net.
- Ahmed Masry, Do Xuan Long, Jia Qing Tan, Shafiq R. Joty, and Enamul Hoque. 2022. [Chartqa: A benchmark for question answering about charts with visual and logical reasoning](#). In *Findings of the Association for Computational Linguistics: ACL 2022, Dublin, Ireland, May 22-27, 2022*, Findings of ACL, pages 2263–2279. Association for Computational Linguistics.
- OpenAI. 2026. Gpt-5.4 thinking system card. <https://openai.com/index/gpt-5-4-thinking-system-card/>. Released March 5, 2026.
- Arka Pal, Deep Karkhanis, Samuel Dooley, Manley Roberts, Siddhartha Naidu, and Colin White. 2024. [Smaug: Fixing failure modes of preference optimization with dpo-positive](#). *CoRR*, abs/2402.13228.
- Ryan Park, Rafael Rafailov, Stefano Ermon, and Chelsea Finn. 2024. [Disentangling length from quality in direct preference optimization](#). In *Findings of the Association for Computational Linguistics, ACL 2024, Bangkok, Thailand and virtual meeting, August 11-16, 2024*, volume ACL 2024 of *Findings of ACL*, pages 4998–5017. Association for Computational Linguistics.
- Renjie Pi, Tianyang Han, Wei Xiong, Jipeng Zhang, Runtao Liu, Rui Pan, and Tong Zhang. 2024. [Strengthening multimodal large language model with bootstrapped preference optimization](#). In *Computer Vision - ECCV 2024 - 18th European Conference, Milan, Italy, September 29-October 4, 2024, Proceedings, Part XXXIII*, volume 15091 of *Lecture Notes in Computer Science*, pages 382–398. Springer.
- 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. 2025. [Judge anything: MLLM as a judge across any modality](#). In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining, V.2, KDD 2025, Toronto ON, Canada, August 3-7, 2025*, pages 5742–5753. ACM.
- 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](#). In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.
- MohammadHossein Rezaei, Robert Vacareanu, Zihao Wang, Clinton Wang, Bing Liu, Yunzhong He, and Afra Feyza Akyurek. 2025. [Online rubrics elicitation from pairwise comparisons](#). *CoRR*, abs/2510.07284.
- Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. 2024. [Deepseekmath: Pushing the limits of mathematical reasoning in open language models](#). *CoRR*, abs/2402.03300.
- Giulio Starace, Oliver Jaffe, Dane Sherburn, James Aung, Jun Shern Chan, Leon Maksin, Rachel Dias, Evan Mays, Benjamin Kinsella, Wyatt Thompson, Johannes Heidecke, Amelia Glaese, and Tejal Patwardhan. 2025. [Paperbench: Evaluating ai’s ability to replicate AI research](#). In *Forty-second International Conference on Machine Learning, ICML 2025, Vancouver, BC, Canada, July 13-19, 2025*, volume 267 of *Proceedings of Machine Learning Research*. PMLR / OpenReview.net.
- Zhiqing Sun, Sheng Shen, Shengcao Cao, Haotian Liu, Chunyuan Li, Yikang Shen, Chuang Gan, Liangyan Gui, Yu-Xiong Wang, Yiming Yang, Kurt Keutzer, and Trevor Darrell. 2024. [Aligning large multimodal models with factually augmented RLHF](#). In *Findings of the Association for Computational Linguistics, ACL 2024, Bangkok, Thailand and virtual meeting, August 11-16, 2024*, volume ACL 2024 of *Findings of ACL*, pages 13088–13110. Association for Computational Linguistics.
- Qwen Team. 2025. [Qwen3-vl technical report](#). *CoRR*, abs/2511.21631.
- Vijay Viswanathan, Yanchao Sun, Shuang Ma, Xiang Kong, Meng Cao, Graham Neubig, and Tongshuang Wu. 2025. [Checklists are better than reward models for aligning language models](#). *CoRR*, abs/2507.18624.
- Fei Wang, Wenxuan Zhou, James Y. Huang, Nan Xu, Sheng Zhang, Hoifung Poon, and Muhao Chen. 2024a. [mdpo: Conditional preference optimization for multimodal large language models](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, EMNLP 2024, Miami, FL, USA, November 12-16, 2024*, pages 8078–8088. Association for Computational Linguistics.
- Weiyun Wang, Zhe Chen, Wenhai Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Jinguo Zhu, Xizhou Zhu, Lewei Lu, Yu Qiao, and Jifeng Dai. 2024b. [Enhancing the reasoning ability of multimodal large language models via mixed preference optimization](#). *CoRR*, abs/2411.10442.
- Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, Zhaokai Wang, Zhe Chen, Hongjie Zhang, Ganlin Yang, Haomin Wang, Qi Wei, Jinhui Yin, Wenhao Li, Erfei Cui, and 56 others. 2025a. [Internvl3.5: Advancing open-source multimodal models in versatility, reasoning, and efficiency](#). *CoRR*, abs/2508.18265.
- Xiaokun Wang, Peiyu Wang, Jiangbo Pei, Wei Shen, Yi Peng, Yunzhuo Hao, Weijie Qiu, Ai Jian, Tianyidan Xie, Xuchen Song, Yang Liu, and Yahui Zhou. 2025b. [Skywork-vl reward: An effective reward model for multimodal understanding and reasoning](#). *CoRR*, abs/2505.07263.
- Xiyao Wang, Jiu hai Chen, Zhaoyang Wang, Yuhang Zhou, Yiyang Zhou, Huaxiu Yao, Tianyi Zhou, Tom Goldstein, Parminder Bhatia, Taha A. Kass-Hout, Furong Huang, and Cao Xiao. 2025c. [Enhancing visual-language modality alignment in large vision language models via self-improvement](#). In *Findings of the Association for Computational Linguistics: NAACL 2025, Albuquerque, New Mexico, USA, April 29 - May 4, 2025*, volume NAACL 2025 of *Findings of ACL*, pages 268–282. Association for Computational Linguistics.
- Zhilin Wang, Jaehun Jung, Ximing Lu, Shizhe Diao, Ellie Evans, Jiaqi Zeng, Pavlo Molchanov, Yejin Choi, Jan Kautz, and Yi Dong. 2025d. [Profbench: Multi-domain rubrics requiring professional knowledge to answer and judge](#). *CoRR*, abs/2510.18941.
- Lipeng Xie, Sen Huang, Zhuo Zhang, Anni Zou, Yunpeng Zhai, Dingchao Ren, Kezun Zhang, Haoyuan Hu, Boyin Liu, Haoran Chen, Zhaoyang Liu, and Bolin Ding. 2025. [Auto-rubric: Learning to extract generalizable criteria for reward modeling](#). *CoRR*, abs/2510.17314.
- Yuxi Xie, Guanzhen Li, Xiao Xu, and Min-Yen Kan. 2024. [V-DPO: mitigating hallucination in large vision language models via vision-guided direct preference optimization](#). In *Findings of the Association for Computational Linguistics: EMNLP 2024, Miami, Florida, USA, November 12-16, 2024*, volume EMNLP 2024 of *Findings of ACL*, pages 13258–13273. Association for Computational Linguistics.
- Shuo Xing, Peiran Li, Yuping Wang, Ruizheng Bai, Yueqi Wang, Chan-Wei Hu, Chengxuan Qian, Huaxiu Yao, and Zhengzhong Tu. 2025. [Re-align: Aligning vision language models via retrieval-augmented direct preference optimization](#). In *Proceedings of the*

- 2025 Conference on Empirical Methods in Natural Language Processing, EMNLP 2025, Suzhou, China, November 4-9, 2025, pages 2379–2397. Association for Computational Linguistics.
- Tianyi Xiong, Yi Ge, Ming Li, Zuolong Zhang, Pranav Kulkarni, Kaishen Wang, Qi He, Zeying Zhu, Chenxi Liu, Ruibo Chen, Tong Zheng, Yanshuo Chen, Xiyao Wang, Renrui Zhang, Wenhui Chen, and Heng Huang. 2025a. [Multi-crit: Benchmarking multi-modal judges on pluralistic criteria-following](#). *CoRR*, abs/2511.21662.
- Tianyi Xiong, Xiyao Wang, Dong Guo, Qinghao Ye, Haoqi Fan, Quanquan Gu, Heng Huang, and Chunyuan Li. 2025b. [Llava-critic: Learning to evaluate multimodal models](#). In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2025, Nashville, TN, USA, June 11-15, 2025*, pages 13618–13628. Computer Vision Foundation / IEEE.
- Ran Xu, Tianci Liu, Zihan Dong, Tony Yu, Ilgee Hong, Carl Yang, Linjun Zhang, Tao Zhao, and Haoyu Wang. 2026. [Alternating reinforcement learning for rubric-based reward modeling in non-verifiable llm post-training](#). *CoRR*, abs/2602.01511.
- Biao Yang, Bin Wen, Boyang Ding, Changyi Liu, Chenglong Chu, Chengru Song, Chongling Rao, Chuan Yi, Da Li, Dunju Zang, Fan Yang, Guorui Zhou, Guowang Zhang, Han Shen, Hao Peng, Haojie Ding, Hao Wang, Haonan Fan, Hengrui Ju, and 42 others. 2025a. [Kwai keye-vl 1.5 technical report](#). *CoRR*, abs/2509.01563.
- Zhihe Yang, Xufang Luo, Dongqi Han, Yunjian Xu, and Dongsheng Li. 2025b. [Mitigating hallucinations in large vision-language models via DPO: on-policy data hold the key](#). In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2025, Nashville, TN, USA, June 11-15, 2025*, pages 10610–10620. Computer Vision Foundation / IEEE.
- Michihiro Yasunaga, Luke Zettlemoyer, and Marjan Ghazvininejad. 2025. [Multimodal rewardbench: Holistic evaluation of reward models for vision language models](#). *CoRR*, abs/2502.14191.
- Chengzhi Yu, Yifan Xu, Yifan Chen, and Wenyi Zhang. 2025a. [Optimizing vlms with on-policy data for effective hallucination mitigation](#). *CoRR*, abs/2512.00706.
- Tianyu Yu, Yuan Yao, Haoye Zhang, Taiwen He, Yifeng Han, Ganqu Cui, Jinyi Hu, Zhiyuan Liu, Hai-Tao Zheng, and Maosong Sun. 2024a. [RLHF-V: towards trustworthy mllms via behavior alignment from fine-grained correctional human feedback](#). In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pages 13807–13816. IEEE.
- Tianyu Yu, Haoye Zhang, Qiming Li, Qixin Xu, Yuan Yao, Da Chen, Xiaoman Lu, Ganqu Cui, Yunkai Dang, Taiwen He, Xiaocheng Feng, Jun Song, Bo Zheng, Zhiyuan Liu, Tat-Seng Chua, and Maosong Sun. 2025b. [RLAIF-V: open-source AI feedback leads to super GPT-4V trustworthiness](#). In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2025, Nashville, TN, USA, June 11-15, 2025*, pages 19985–19995. Computer Vision Foundation / IEEE.
- Ya-Qi Yu, Minghui Liao, Jiwen Zhang, and Jihao Wu. 2024b. [Texthawk2: A large vision-language model excels in bilingual OCR and grounding with 16x fewer tokens](#). *CoRR*, abs/2410.05261.
- Zihao Yue, Zhenru Lin, Yifan Song, Weikun Wang, Shuhuai Ren, Shuhao Gu, Shicheng Li, Peidian Li, Liang Zhao, Lei Li, Kainan Bao, Hao Tian, Hailin Zhang, Xiao-Gang Wang, Dawei Zhu, Cici, Chenhong He, Bowen Ye, Bowen Shen, and 54 others. 2025. [Mimo-vl technical report](#). *CoRR*, abs/2506.03569.
- Yuhang Zang, Xiaoyi Dong, Pan Zhang, Yuhang Cao, Ziyu Liu, Shengyuan Ding, Shenxi Wu, Yubo Ma, Haodong Duan, Wenwei Zhang, Kai Chen, Dahua Lin, and Jiaqi Wang. 2025. [Internlm-xcomposer2.5-reward: A simple yet effective multi-modal reward model](#). In *Findings of the Association for Computational Linguistics, ACL 2025, Vienna, Austria, July 27 - August 1, 2025*, volume ACL 2025 of *Findings of ACL*, pages 6547–6563. Association for Computational Linguistics.
- Di Zhang, Jingdi Lei, Junxian Li, Xunzhi Wang, Yujie Liu, Zonglin Yang, Jiatong Li, Weida Wang, Suorong Yang, Jianbo Wu, Peng Ye, Wanli Ouyang, and Dongzhan Zhou. 2025a. [Critic-v: VLM critics help catch VLM errors in multimodal reasoning](#). In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2025, Nashville, TN, USA, June 11-15, 2025*, pages 9050–9061. Computer Vision Foundation / IEEE.
- Yifan 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, Tingting Gao, Zhang Zhang, Fan Yang, Di Zhang, Liang Wang, and Rong Jin. 2025b. [MM-RLHF: the next step forward in multimodal LLM alignment](#). In *Forty-second International Conference on Machine Learning, ICML 2025, Vancouver, BC, Canada, July 13-19, 2025*, volume 267 of *Proceedings of Machine Learning Research*. PMLR / OpenReview.net.
- Zhiyuan Zhao, Bin Wang, Linke Ouyang, Xiaoyi Dong, Jiaqi Wang, and Conghui He. 2025. [Beyond multimodal hallucinations: Enhancing vlms through hallucination-aware direct preference optimization](#). In *IEEE International Conference on Multimedia and Expo, ICME 2025, Nantes, France, June 30 - July 4, 2025*, pages 1–6. IEEE.
- Yang Zhou, Sunzhu Li, Shunyu Liu, Wenkai Fang, Jiale Zhao, Jingwen Yang, Jianwei Lv, Kongcheng Zhang, Yihe Zhou, Hengtong Lu, Wei Chen, Yan Xie, and

Mingli Song. 2025. [Breaking the exploration bottleneck: Rubric-scaffolded reinforcement learning for general LLM reasoning](#). *CoRR*, abs/2508.16949.

Yiyang Zhou, Chenhang Cui, Rafael Rafailov, Chelsea Finn, and Huaxiu Yao. 2024. [Aligning modalities in vision large language models via preference fine-tuning](#). *CoRR*, abs/2402.11411.

Ke Zhu, Liang Zhao, Zheng Ge, and Xiangyu Zhang. 2024. [Self-supervised visual preference alignment](#). In *Proceedings of the 32nd ACM International Conference on Multimedia, MM 2024, Melbourne, VIC, Australia, 28 October 2024 - 1 November 2024*, pages 291–300. ACM.

A Prompt Templates for Rubric Construction and Scoring

This appendix presents the complete system prompts used across the stages of rubric generation, rubric aggregation, and reward modeling, alongside the JSON schemas for their respective outputs. Each prompt is delivered as the system message to the reasoning model; the corresponding user message supplies the image and question, together with any stage-specific auxiliary inputs (ground-truth annotations or candidate rubric lists).

A.1 Output JSON Schemas

Both rubric generation and response scoring produce structured JSON outputs, which facilitate reliable parsing and enable fine-grained programmatic aggregation and filtering.

Rubric Schema. The checklist-style rubric is a JSON object with two top-level arrays: *essential* and *additional*, corresponding to the two criterion categories defined in Section 4.2. Each element in either array is a triplet:

- **criterion:** a concrete, observable assertion that the judge can verify directly from the image, question, and candidate response.
- **reference:** the ground-truth answer strictly grounded in image facts or common knowledge.
- **weight:** a three-level integer encoding the criterion’s importance—**1** (Auxiliary: supplementary information), **2** (Important: noticeably affects the user experience), **3** (Key: critical elements where any omission or deviation constitutes a definitive error).

```
{
  "essential": [
    {
      "criterion": "string",
      "reference": "string",
      "weight": 1 | 2 | 3
    }
  ],
  "additional": [
    ...
  ]
}
```

Scoring Schema. The scoring output is a JSON array with one element per criterion, preserving the same ordering as the input rubric. Each element contains three fields:

- **criterion:** the assertion copied verbatim from

the input rubric, serving as an explicit confirmation that the judge is addressing the correct item.

- **rationale**: a brief (1–2 sentence) explanation of the judgment, providing transparency and enabling post-hoc error analysis.
- **credit**: a three-level score—**0** (incorrect or missing), **0.5** (partially correct with minor errors), **1** (fully correct or semantically equivalent). Critically, *credit* is a *qualitative correctness grade* independent of *weight*; the weighted aggregation across all scored criteria is performed externally (Eq. 8).

```
[
  {
    "criterion": "string",
    "rationale": "string",
    "credit": 0 | 0.5 | 1
  }
]
```

A.2 Rubric Generation

The generation prompt instructs the reasoning model to construct an instance-specific checklist-style rubric for a given image-question pair, following the principles described in Section 4.2:

Expert-Grounded Generation. When trustworthy ground-truth annotations are available, the prompt includes a *dual verification* step: before finalizing the rubric, the model must confirm that the provided ground-truth fully satisfies all *essential* criteria, ensuring that no correct response would be penalized by an erroneous check item.

System Prompt: Expert-Grounded Rubric Generation

You are a multimodal evaluation expert. Given the user question and associated image, construct an instance-specific Checklist-Style Rubric for evaluating the accuracy of model responses. The criteria in this Checklist serve as the *sole evaluation standard* for the reward model.

Construction Principles

- **Atomic**: Each check item targets exactly one key point or atomic sub-question within the query.
- **Comprehensive**: The combined set of items covers all critical dimensions of the user’s question.
- **Precise**: Exclude redundant checks and checks unrelated to the question.
- **Objective**: Ground assessments in image facts or logical truths; avoid subjective uncertainty.

Field Definitions

Two categories of check items:

- **essential**: Core information prioritized by the

query. Satisfying these is a prerequisite for a conceptually sound response regardless of its verbosity.

- **additional**: Relevant image facts, supplementary knowledge, or intermediate steps required to derive the answer. Optional, the list might be empty.

Fields per item:

- **criterion**: A concrete and observable assertion to be verified.
- **reference**: Ground-truth fact derived from the image and common knowledge.
- **weight**: A three-level integer quantifying the criterion’s importance ranging from 1 (Auxiliary: supplementary information) through 2 (Important: noticeable impact on the user experience) to 3 (Key: critical elements where any omission or deviation constitutes a definitive error).

Think from a scoring perspective: do not double-count and avoid over-decomposing key points.

Dual Verification

Before finalizing the Checklist, verify that the provided reference answer satisfies all *essential* items. Any unsatisfied item indicates a construction error—the item is either incorrect or superfluous.

Output Format

{Structured JSON following the rubric schema in Appendix A.1.}

Answer-Agnostic Generation. For datasets with noisy or absent annotations, the *dual verification* step is omitted, preventing the model from anchoring on a potentially unreliable reference answer.

System Prompt: Answer-Agnostic Rubric Generation

You are a multimodal evaluation expert. Given the user question and associated image, construct an instance-specific Checklist-Style Rubric for evaluating the accuracy of model responses. The criteria in this Checklist serve as the *sole evaluation standard* for the reward model.

Construction Principles

- **Atomic**: Each check item targets exactly one key point or atomic sub-question within the query.
- **Comprehensive**: The combined set of items covers all critical dimensions of the user’s question.
- **Precise**: Exclude redundant checks and checks unrelated to the question.
- **Objective**: Ground assessments in image facts or logical truths; avoid subjective uncertainty.

Field Definitions

Two categories of check items:

- **essential**: Core information prioritized by the query. Satisfying these is a prerequisite for a conceptually sound response regardless of its verbosity.
- **additional**: Relevant image facts, supplementary knowledge, or intermediate steps required to derive the answer. Optional, the list might be empty.

Fields per item:

- **criterion**: A concrete and observable assertion to be verified.
- **reference**: Ground-truth fact derived from the image and common knowledge.
- **weight**: A three-level integer quantifying the criterion’s importance ranging from 1 (Auxiliary: sup-

plementary information) through 2 (Important: noticeable impact on the user experience) to 3 (Key: critical elements where any omission or deviation constitutes a definitive error).

Think from a scoring perspective: do not double-count and avoid over-decomposing key points.

Output Format

{Structured JSON following the rubric schema in Appendix A.1.}

A.3 Rubric Aggregation

After independent rubric generation by multiple models (Section 4.2), an aggregation prompt merges the candidate checklists into a single unified rubric. The model first applies the same four construction principles as in rubric generation, then executes the additional aggregation instructions below to apply majority-vote filtering, deduplicate overlapping check items, and verify the correctness of all references.

System Prompt: Rubric Aggregation

You are a multimodal evaluation expert. Given the user question and associated image, construct an instance-specific Checklist-Style Rubric for evaluating the accuracy of model responses. The criteria in this Checklist serve as the *sole evaluation standard* for the reward model.

Construction Principles

- **Atomic:** Each check item targets exactly one key point or atomic sub-question within the query.
- **Comprehensive:** The combined set of items covers all critical dimensions of the user's question.
- **Precise:** Exclude redundant checks and checks unrelated to the question.
- **Objective:** Ground assessments in image facts or logical truths; avoid subjective uncertainty.

Field Definitions

Two categories of check items:

- **essential:** Core information prioritized by the query. Satisfying these is a prerequisite for a conceptually sound response regardless of its verbosity.
- **additional:** Relevant image facts, supplementary knowledge, or intermediate steps required to derive the answer. Optional, the list might be empty.

Fields per item:

- **criterion:** A concrete and observable assertion to be verified.
- **reference:** Ground-truth fact derived from the image and common knowledge.
- **weight:** A three-level integer quantifying the criterion's importance ranging from 1 (Auxiliary: supplementary information) through 2 (Important: noticeable impact on the user experience) to 3 (Key: critical elements where any omission or deviation constitutes a definitive error).

Think from a scoring perspective: do not double-count and avoid over-decomposing key points.

Checklist Aggregation

Do not construct the Checklist from scratch. Multiple models have each independently generated a candidate

Checklist for the same question; your task is to merge them into a single unified rubric. Based on the construction principles, ensure that:

- All retained check items are necessary. Specifically:
 - Retain only items that the majority of candidates agree should be checked.
 - Remove redundant or duplicate checks.
 - Remove checks unrelated to the question.
- Do not over-decompose check items.
- All reference values are correct.

Output Format

{Structured JSON following the rubric schema in Appendix A.1.}

A.4 Response Scoring

The scoring prompt instructs the judge model to evaluate a candidate response against the finalized rubric, assigning a credit score to each criterion independently on a three-point scale.

System Prompt: Response Scoring

You are a multimodal evaluation expert. Given the user question, associated image, and a Checklist-Style Rubric, evaluate the accuracy of the model response. Make sure that the criteria in the Checklist are used as the *sole evaluation standard*.

Checklist Fields

Two categories of check items:

- **essential:** Core information prioritized by the query. Satisfying these is a prerequisite for a conceptually sound response regardless of its verbosity.
- **additional:** Relevant image facts, supplementary knowledge, or intermediate steps required to derive the answer. Optional, the list might be empty.

Fields per item:

- **criterion:** A concrete and observable assertion to be verified.
- **reference:** Ground-truth fact derived from the image and common knowledge.

Scoring Fields

For each criterion, output:

- **criterion:** The assertion to be verified (*copied verbatim from input*).
- **rationale:** Core reasoning for the judgment (1–2 sentences).
- **credit:** A three-level score ranging from 0 (No Credit: incorrect or missing) through 0.5 (Partial Credit: partially correct with minor errors) to 1 (Full Credit: fully correct or semantically equivalent).

Output Format

{Structured JSON following the scoring schema in Appendix A.1.}

B Rubric Quality and Human Audit

B.1 Human Audit Protocol

We conduct a category-stratified diagnostic audit of the data-construction pipeline. Twelve multimodal data and algorithm engineers participate, with each auditor responsible for one or two data categories.

Table 5: Criterion-level agreement accuracy (%) against Gemini 3.0 Pro pseudo-references.

Decoding	Single	Median@3	Median@9	Mode@9
Qwen3-VL-30B-A3B-Instruct				
$T = 1.0$	90.77	91.48	–	–
$T = 0.7, p = 0.8, k = 20$	91.33	92.17	–	–
Greedy	91.34	–	–	–
Qwen3-VL-32B-Instruct				
$T = 1.0$	92.45	93.34	93.47	93.87
$T = 0.7, p = 0.8, k = 20$	92.63	93.37	93.42	93.51
Greedy	92.83	–	–	–

Rather than correcting isolated erroneous instances, the auditors identify recurring failure modes attributable to the pipeline so that the corresponding fixes can be applied consistently across other instances and target policies. For example, observations of overlapping criteria motivated the *Precise* principle, while cases in which detailed but incorrect responses were favored over concise correct ones motivated the separation of essential and additional criteria.

B.2 Criterion-Level Scoring Stability

We assess scorer stability on a sampled subset of the scaling-training pool. For each rubric criterion, Gemini 3.0 Pro scores the response three times, and the median score is used as a pseudo-reference. We then measure criterion-level agreement accuracy between this pseudo-reference and scores produced by Qwen3-VL-30B-A3B-Instruct or Qwen3-VL-32B-Instruct.

Table 5 reports our three-pass setting and a nine-pass stability check. Three-pass median aggregation improves agreement by 0.70–0.86 points. For Qwen3-VL-32B, increasing to nine passes adds only 0.12 and 0.05 points to the median results. Nine-pass mode adds a further 0.40 and 0.09 points, but triples the number of scoring calls relative to our deployed setting. We therefore use three-pass median by default.

B.3 Safeguards Against Noise

The pipeline addresses distinct sources of noise at different stages. Multi-model rubric drafting and aggregation target recurring construction errors, while repeated scoring and median aggregation reduce sensitivity to stochastic judge outputs. During pair mining, the qualification rules prevent strong scores on additional criteria from compensating for failures on essential criteria. The reward-margin constraint further removes near-tied pairs, whose ordering is particularly sensitive to scoring vari-

ance, and max-margin selection prioritizes clearer comparisons. Diversity selection instead limits repeated use of highly similar responses.

C Scaling Evaluation

The comprehensive in-house benchmark is used to assess scaling to a broader set of multilingual multimodal tasks under response-style constraints. This section documents the benchmark composition and response-style constraints underlying the scaling-validation results in Table 3.

C.1 Benchmark

The comprehensive benchmark contains 3,197 examples spanning four broad categories and 21 sub-categories, as reported in Table 6.

C.2 Style Constraints

The preference-optimization setting aims to improve task accuracy subject to three additional response-style constraints: (1) responses use conversational plain text without Markdown formatting; (2) the mean response length does not exceed 150 tokens; and (3) responses are direct and answer-first, without reflective or meta-discursive passages. Representative before-and-after examples are provided in the preference-optimization case study in Section E.

D Case Study of Reward Modeling

The following cases are sampled from the scaling-training data and illustrate how rubric scoring resolves response-level preferences at the criterion level. Each checklist separates essential and additional criteria and reports the rationale and credit assigned to both responses.

Count	
	
Question: How many planes are visible? Response A: Three planes are visible. Response B: Three planes are visible. The one on top is an F-22 stealth fighter, the one in the middle is a P-51 Mustang propeller plane, and the one below is an F-16 fighter jet. Checklist:	
Type: Essential	Weight: 3

Table 6: Composition of the comprehensive in-house benchmark.

Category	Subcategory	Count	Within cat. (%)	Overall (%)
Perception (<i>N</i> = 503)	Fine-Grained Object & Attribute Recognition	162	32.2	5.1
	Scene & Event Recognition	129	25.6	4.0
	Optical Character & Symbol Recognition	105	20.9	3.3
	Imaging Modality & Photographic Attribute Perception	56	11.1	1.8
	Image Aesthetics & Quality Assessment	51	10.1	1.6
Comprehension (<i>N</i> = 614)	Key Information Extraction (Cards & Receipts)	230	37.5	7.2
	Multimodal & Scene-Text Translation	120	19.5	3.8
	Visual Semantic Classification (Entity/Topic/Intent)	120	19.5	3.8
	Visually-Rich Document & Multi-Image VQA	90	14.7	2.8
	Key Information Extraction (Open-Domain Documents)	54	8.8	1.7
Knowledge (<i>N</i> = 449)	Consumer Product & Commerce Knowledge	115	25.6	3.6
	Culture, Tourism & Entertainment Knowledge	112	24.9	3.5
	Natural Science Knowledge	107	23.8	3.3
	Humanities & Social Science Knowledge	69	15.4	2.2
	Biomedical & Applied Technology Knowledge	46	10.2	1.4
Reasoning (<i>N</i> = 1,631)	Academic Mathematical Reasoning	425	26.1	13.3
	Real-World Quantitative Reasoning	340	20.8	10.6
	Commonsense & Causal Reasoning	309	18.9	9.7
	Abstract Visual Reasoning (Induction & Analogy)	243	14.9	7.6
	Spatio-Temporal Reasoning	177	10.9	5.5
	Deductive Reasoning & Multi-Step Planning	137	8.4	4.3
Total		3,197	—	100.0

Criterion: Accurately state the total number of plane visible in the image.

Reference: 3

Rationale for A: The model response correctly identifies that there are three planes visible in the image, which matches the reference. **Credit: 1**

Rationale for B: The model correctly states that there are three planes visible in the image, which matches the reference. **Credit: 1**

Type: Additional

Weight: 1

Criterion: Identify the aircraft model in the top-left.

Reference: F-22 Raptor

Rationale for A: The model response does not provide any identification of the aircraft models, so this criterion is not addressed. **Credit: 0**

Rationale for B: The model identifies the top-left aircraft as an F-22 stealth fighter, which is correct and matches the reference. **Credit: 1**

Type: Additional

Weight: 1

Criterion: Identify the aircraft model in the middle.

Reference: P-51 Mustang

Rationale for A: The model response does not provide any identification of the aircraft models, so this criterion is not addressed. **Credit: 0**

Rationale for B: The model correctly identifies the middle aircraft as a P-51 Mustang, which is accurate and aligns with the reference. **Credit: 1**

Type: Additional

Weight: 1

Criterion: Identify the aircraft model at the bottom.

Reference: F-16 Fighting Falcon

Rationale for A: The model response does not provide any identification of the aircraft models, so this criterion is not addressed. **Credit: 0**

Rationale for B: The model correctly identifies the bottom aircraft as an F-16 fighter jet, which is accurate and matches the reference. **Credit: 1**

VQA



Question: Is the door open?

Response A: Yes, the door is open. The image shows the white fence gate in an open position, providing access to the courtyard behind.

Response B: Yes, the gate in the center of the image is open. The black door on the right is closed.

Checklist:

Type: Essential

Weight: 3

Criterion: Determine the status of the white fence gate in the picture.

Reference: The white fence gate in the picture is open.

Rationale for A: The model correctly identifies that the white fence gate is open, which aligns with the reference. The gate is visibly ajar, allowing access to the courtyard. **Credit: 1**

Rationale for B: The model correctly identifies that the white gate in the center of the image is open, which matches the reference. **Credit: 1**

Type: Essential

Weight: 3

Criterion: Determine the status of the black door on the building on the right.

Reference: The black door in the picture is closed.

Rationale for A: The model fails to address the black door on the right, which is closed. The response incorrectly implies the door is open, which is not supported by the image. **Credit: 0**

Rationale for B: The model correctly states that the black door on the right is closed, which is consistent with the reference. **Credit: 1**

Type: Additional

Weight: 1

Criterion: Determine the status of the door on the white building on the left.

Reference: The door on the left in the picture is closed.

Rationale for A: The model does not mention the door on the left building, which is closed. While this is an additional criterion, the

model's failure to address the primary open door (the gate) is the main issue. **Credit: 0**

Rationale for B: The model does not address the status of the door on the left building, which is an additional criterion. However, since the question only asks about 'the door' and the most prominent door is the black one, the response is still considered accurate. **Credit: 0.5**

Knowledge



Question: Which architect designed this historic house museum?

Response A: **Bridgeman** designed this historic house; he is a famous British master architect whose works integrate classical styles.

Response B: This historic house is **Blenheim Palace**, designed by English architect **Sir John Vanbrugh**. He incorporated **Baroque** styles to make the palace appear grand and magnificent.

Checklist:

Type: Essential

Weight: 3

Criterion: Accurately identify the name of the building in the picture.

Reference: Blenheim Palace, also frequently called the Churchill Manor.

Rationale for A: The model response does not identify the building as Blenheim Palace, which is the correct name of the structure shown in the image. Instead, it incorrectly refers to 'Bridgeman,' which is not the name of the building. **Credit: 0**

Rationale for B: The model correctly identifies the building as Blenheim Palace, which aligns with the reference. **Credit: 1**

Type: Essential

Weight: 3

Criterion: Accurately state the name of the primary architect who designed the building.

Reference: Sir John Vanbrugh.

Rationale for A: The model response incorrectly identifies the architect as 'Bridgeman,' which is factually incorrect. The actual primary architect of Blenheim Palace is Sir John Vanbrugh. The response fails to provide the correct name. **Credit: 0**

Rationale for B: The model correctly states that Sir John Vanbrugh was the primary architect, matching the reference. **Credit: 1**

Type: Additional

Weight: 1

Criterion: Mention another architect who made significant contributions to the design of the building.

Reference: Nicholas Hawksmoor.

Rationale for A: The model response does not mention Nicholas Hawksmoor, who was a significant contributor to the design of Blenheim Palace. This information is missing from the response. **Credit: 0**

Rationale for B: The model does not mention Nicholas Hawksmoor, who is noted in the reference as a significant contributor, so this criterion is not met. **Credit: 0**

Type: Additional

Weight: 1

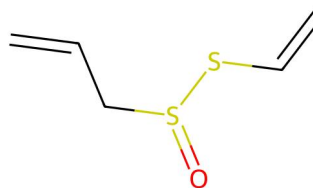
Criterion: Mention the architectural style of the building.

Reference: English Baroque style.

Rationale for A: The model response does not mention the architectural style of the building, which is English Baroque. This information is absent from the response. **Credit: 0**

Rationale for B: The model correctly identifies the architectural style as Baroque, which is consistent with the reference. **Credit: 1**

Molecular



Question: What is the molecular formula of alliin?

Response A: **The structure you provided is alliin**, and its molecular formula is $C_6H_{10}S_2$. It contains two sulfur atoms and two allyl groups, which is why it's also known as **diallyl disulfide**. It is the primary source of garlic's pungent flavor.

Response B: **Based on the image you provided**, the chemical formula of alliin (**chemical name: diallyl thiosulfinate**) is: $C_6H_{10}OS_2$.

Checklist:

Type: Essential

Weight: 3

Criterion: Correctly answer the text-based question and identify the molecular formula of Alliin.

Reference: $C_6H_{10}OS_2$

Rationale for A: The model incorrectly identifies the molecular formula as $C_6H_{10}S_2$, which is inaccurate. The correct molecular formula for alliin is $C_6H_{10}OS_2$. Additionally, the model misidentifies the compound in the image as alliin, which is incorrect. **Credit: 0**

Rationale for B: The model correctly identifies the molecular formula of alliin as $C_6H_{10}OS_2$, which matches the reference. **Credit: 1**

Type: Additional

Weight: 2

Criterion: Clearly state that the molecule in the image is not the same substance as the Alliin mentioned in the question, and provide the correct molecular formula for the molecule in the image.

Reference: The molecular structure in the image is different from that of alliin; its molecular formula is $C_5H_8OS_2$.

Rationale for A: The model fails to recognize that the molecule in the image is not alliin. It incorrectly states that the provided structure is alliin and does not provide the correct molecular formula for the actual molecule shown ($C_5H_8OS_2$). **Credit: 0**

Rationale for B: The model fails to acknowledge that the molecule in the provided image is structurally different from alliin and does not provide the correct molecular formula ($C_5H_8OS_2$) for the structure shown. **Credit: 0**

Type: Additional

Weight: 1

Criterion: Provide the chemical names for Alliin and the molecule in the image respectively.

Reference: The chemical name of Alliin is diallyl thiosulfinate. The chemical name of the molecule in the image is S-allyl ethenesulfinothioate (or S-allyl vinyl thiosulfinate).

Rationale for A: The model does not provide the chemical name for alliin or the molecule in the image. It incorrectly identifies the image as alliin and calls it diallyl disulfide, which is a different compound. **Credit: 0**

Rationale for B: The model provides the correct chemical name for alliin ($C_6H_{10}OS_2$) but does not provide the name for the molecule in the image. **Credit: 0.5**

Type: Additional

Weight: 1

Criterion: Briefly explain the core structural difference between Alliin and the molecule in the image.

Reference: Alliin's structure consists of two allyl groups attached to the central thiosulfinate group ($-S(O)S-$), whereas the molecule in the image consists of one allyl group and one vinyl group attached to that same central structure.

Rationale for A: The model does not explain any structural differences because it incorrectly assumes the image shows alliin. It fails to recognize the structural difference between alliin (two allyl groups) and the molecule in the image (one allyl and one vinyl group). **Credit: 0**

Rationale for B: The model does not explain the structural difference between alliin and the molecule in the image. **Credit: 0**

TabMWP

	Red	Green
Big	2	1
Small	5	3

Question: If Romero picks a marble at random, what is the probability that it is not red?

Response A: There are a total of $2 + 1 + 5 + 3 = 11$ marbles, among which there are $2 + 1 = 3$ red ones. There are $11 - 3 = 8$ non-red marbles. Therefore, the probability of it being non-red is $8/11$.

Response B: The total number of marbles is $2 + 1 + 5 + 3 = 11$. The non-red ones are green, and there are $1 + 3 = 4$. Therefore, the probability of it being non-red is $4/11$.

Checklist:

Type: Essential **Weight:** 3
Criterion: Accurately provide the probability of randomly picking a non-red marble.
Reference: $4/11$

Rationale for A: The model incorrectly calculates the number of red marbles as 3 ($2 + 1$) instead of 7 ($2 + 5$), leading to an incorrect count of non-red marbles (8 instead of 4) and a final probability of $8/11$ instead of the correct $4/11$. **Credit: 0**

Rationale for B: The model correctly calculates the probability of picking a non-red marble as $4/11$, which matches the reference value. **Credit: 1**

Type: Additional **Weight:** 1
Criterion: Correctly calculate the total number of all marbles.
Reference: According to the table, the total number of marbles is $2 + 1 + 5 + 3 = 11$.

Rationale for A: The model correctly calculates the total number of marbles as 11 ($2 + 1 + 5 + 3$), which matches the reference. **Credit: 1**

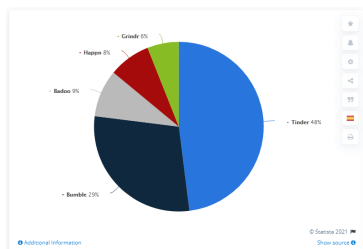
Rationale for B: The model correctly sums the values in the table ($2 + 1 + 5 + 3$) to find the total number of marbles as 11, which is accurate. **Credit: 1**

Type: Additional **Weight:** 1
Criterion: Correctly calculate the total number of non-red (i.e., green) marbles.
Reference: According to the table, non-red marbles are the green marbles, and the count is $1 + 3 = 4$.

Rationale for A: The model incorrectly identifies non-red marbles as 8, but the correct count is 4 (1 green big + 3 green small). The model's error stems from misidentifying the red marbles. **Credit: 0**

Rationale for B: The model correctly identifies the green marbles as the non-red ones and sums their counts ($1 + 3$) to get 4, which is accurate. **Credit: 1**

ChartQA



Question: Which is the most popular dating app in Mexico in 2021?

Response A: In 2021, the most popular dating app in Mexico was Tinder, accounting for 48% of the market. It was followed by Bumble (29%) and Badoo (9%).

Response B: Based on the pie chart provided, the app with the largest share is Tinder at 48%.

However, please note that the flag icon on the right side of the image is the flag of Spain, which suggests this data is for Spain, not Mexico.

Checklist:

Type: Essential **Weight:** 3
Criterion: Based on the chart, clearly identify the most popular dating app and its market share.

Reference: According to the chart, the most popular dating app is Tinder, with a market share of 48%.

Rationale for A: The model correctly identifies Tinder as the most popular app with 48% market share, which is directly supported by the chart data. **Credit: 1**

Rationale for B: The model correctly identifies Tinder as the most popular app with a 48% share, which is the largest segment in the pie chart. **Credit: 1**

Type: Additional **Weight:** 1
Criterion: Point out that the geographical scope of the chart data may not match the user's question and provide a reason.

Reference: The user's question concerns Mexico, but there is a Spanish flag icon in the toolbar on the right side of the chart, which may indicate the data pertains to Spain rather than Mexico.

Rationale for A: The model fails to address the discrepancy between the question about Mexico and the Spanish flag icon in the chart, which suggests the data might be for Spain, not Mexico. **Credit: 0**

Rationale for B: The model correctly observes the Spanish flag icon and correctly infers that the data is likely for Spain, not Mexico, which is a key point of discrepancy. **Credit: 1**

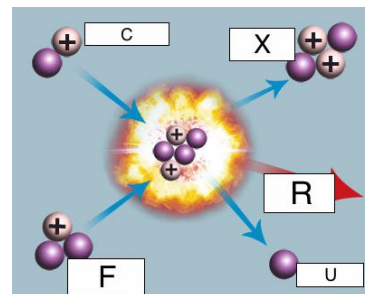
Type: Additional **Weight:** 1
Criterion: List the other dating apps in the chart and their shares in descending order of market share.

Reference: The other apps in the chart and their market shares are: Bumble (29%), Badoo (9%), Happn (8%), and Grindr (6%).

Rationale for A: The model lists Bumble (29%) and Badoo (9%), which are correct, but omits the other apps (Happn, Grindr) and does not present them in descending order. **Credit: 0**

Rationale for B: The model does not list the other apps or their market shares, which is a missing piece of information from the checklist. **Credit: 0**

TQA



Question: Which label refers to the neutron?

Response A: In the image, C is a positively charged particle, which is a proton. X is a neutral particle, uncharged, and is emitted from the nuclear reaction. Neutrons are uncharged and are common products of nuclear reactions. Therefore, X refers to the neutron.

Response B: The image shows the process of nuclear fusion, with atomic nuclei at the center surrounded by positively charged particles (protons) and neutral particles. Label U points to an uncharged particle, which fits the characteristics of a neutron.

Checklist:

Type: Essential **Weight:** 3
Criterion: Correctly select the letter refers to the neutron.
Reference: U

Rationale for A: The model incorrectly identifies label X as the neutron. According to the checklist, the correct label for the neutron is U. The model's answer is factually incorrect. **Credit: 0**

Rationale for B: The model correctly identifies that label U refers to the neutron, which is the uncharged particle shown in the image. **Credit: 1**

Type: Additional **Weight:** 2
Criterion: Explain why label U represents a neutron based on the image information.

Reference: In the image, spheres with a "+" sign are protons, and the uncharged purple spheres are neutrons. Label U points to a single, uncharged purple sphere, therefore representing a neutron.

Rationale for A: The model does not address this criterion. It fails to explain why U represents a neutron, instead providing a flawed explanation for X. This criterion is completely missed. **Credit: 0**

Rationale for B: The model correctly explains that U points to an uncharged particle, which is the defining characteristic of a neutron, consistent with the image's depiction. **Credit: 1**

Type: Additional

Weight: 1

Criterion: Identify the physical process shown in the figure.

Reference: The image shows a nuclear fusion reaction. Specifically, a deuterium nucleus (C) and a tritium nucleus (F) fuse to produce a helium nucleus (X) and a neutron (U), while releasing energy (R).

Rationale for A: The model does not identify the physical process shown in the figure. This criterion is not addressed at all. **Credit: 0**

Rationale for B: The model correctly identifies the process as nuclear fusion, which is the correct interpretation of the image showing nuclei combining and releasing energy. **Credit: 1**

Type: Additional

Weight: 1

Criterion: Explain why the other options are incorrect by identifying the particles represented by labels X, F, and C.

Reference: The other options are incorrect. C represents a deuterium nucleus (1 proton, 1 neutron), F represents a tritium nucleus (1 proton, 2 neutrons), and X represents a helium nucleus (2 protons, 2 neutrons). They are all atomic nuclei composed of multiple particles, rather than a single neutron.

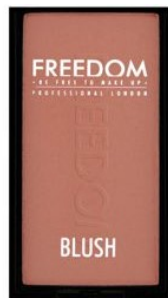
Rationale for A: The model incorrectly identifies X as the neutron, so it cannot provide a correct explanation for why the other labels are incorrect. It fails to identify C and F correctly, and its explanation for X is wrong. This criterion is not addressed correctly. **Credit: 0**

Rationale for B: The model does not address the other labels (X, F, C) or explain why they are incorrect, which is a significant omission from the checklist. **Credit: 0**

E Preference Optimization Cases

The following cases are sampled from the comprehensive in-house benchmark and compare outputs before and after rDPO.

Misleading



Question: Based on the information on the cover, which series does this book likely belong to?

Before: Based on the information on the cover, this book may belong to the "FREEDOM" series.

After: Based on the information in the picture, this is likely not a book, but a makeup product—a blush.

It belongs to the "FREEDOM" series, and the brand is Professional Formula. The packaging says "BE FREE TO MAKE UP," emphasizing the concept of makeup freedom. The shade of this blush is "BLUSH," a warm nude pink.

Fine-Grained Perception



Question: Is there a cobblestone road surface?

Before: There's no cobblestone path in the picture; the ground is dirt and grass, with green mats nearby. It looks like an outdoor children's play area.

After: Yes, there is a cobblestone path on the right side of the picture. This path extends from the lower right corner of the image into the distance, with grassy areas and some tourists strolling along it.

Knowledge



Question: Which movie is this from?

Before: This is a scene from the movie "The Wolf of Wall Street", in which Leonardo DiCaprio plays the protagonist, Jordan Belfort.

After: This is a scene from the film "Catch Me If You Can", starring Leonardo DiCaprio. Directed by Steven Spielberg and released in 2002, the film is an adaptation of Frank Abagnale's autobiography.

Count



Question: If the number of bottles doubles, how many bottles will you see?

Before: There were originally 4 bottles; doubling that number results in 8.

Count the bottles in the picture: 3 on the left, 2 on the right, for a total of 5.

Therefore, if the number of bottles is doubled, you will see 10 bottles.

After: In the image, the following bottles are present:

- 3 small bottles on the left
 - 1 bottle with blue cap in the middle
 - 2 bottles on the right (one tall, one short)
 - 1 bottle behind the fruit basket, partially visible
- That's 7.

So if bottle count becomes 2 times, then $7 * 2 = 14$. Therefore, you would see 14 bottles.

OCR



Question: 最下面的那个蓝色牌子写了啥
What does the blue sign at the very bottom say?

Before: 牌子上写着“请勿在花丛中踩踏”。
The sign says: "Do not trample on the flowers."

After: 最下面的蓝色牌子上写着：“请爱护花草树木”
The blue sign at the bottom says: "Please care for the flowers, grass, and trees."

This bus is Route 16 from Guilin, Guangxi. The body displays "Gui C 03323D" and belongs to the Guilin Bus Group. It has a "Pure Electric, 0 Emissions" label, indicating it is an environmentally friendly model. Based on the route and the license plate, it can be determined that this is part of the Guilin city bus system.

Cuisine



Question: 这个小龙虾开过背了吗
Has this crawfish been deveined (back-cut)?

Before: 看这盘小龙虾，外壳红亮，个头饱满，确实挺诱人的。不过从图片上看，这些小龙虾的背部没有明显的开刀痕迹，应该还是完整的。不过有些店家会提前处理好，所以也有可能是特意没开背。您想吃的时候可以问问服务员哦。

Looking at this plate of crawfish, the shells are bright red and the size is plump, which looks very tempting. However, from the image, there are no obvious cut marks on the backs of these crawfish; they appear to be intact. Some shops might process them beforehand, so it's possible they were intentionally not cut along the back. You can ask the waiter when you are ready to eat.

After: 从图片来看，这些小龙虾的虾背是打开的。可以看到很多小龙虾的背部被切开，露出了里面的虾线和虾肉，这是烹饪前常见的处理方式，方便入味和食用。

另外，它们的颜色非常红亮，酱汁浓郁，看起来是经过爆炒或香辣等烹饪手法处理过的。这种做法通常会先对小龙虾进行开背处理，再进行后续的烹饪步骤。所以可以判断，这些小龙虾已经开过背了。

From the image, the backs of these crawfish are open. You can see that the backs of many crawfish have been cut, exposing the vein and the meat inside. This is a common processing method before cooking to make it easier to season and eat. Additionally, their color is very bright red and the sauce is rich, appearing to have been stir-fried or cooked with spicy seasoning. This preparation method usually involves deveining the crawfish before the subsequent cooking steps. Therefore, it can be determined that these crawfish have been back-cut.

Implicit Reasoning



Question: 帮我记一下我的车位号
Help me remember my parking space number.

Before: 好的，您的车位号是006。下次停车时记得找这个号码哦，我也会帮您记住的。
Sure, your parking space number is 006. Remember to look for this number next time you park, and I will help you remember it as well.

After: 好的，您的车位号是007。旁边还有个006车位，您停在了绿色的标线区域内，位置很清晰。出口方向就在右边柱子上，写着“出口EXLL”和箭头哦。
Sure, your parking space number is 007. There is a space 006 next to it; you are parked within the green marked area, and the location is very clear. The exit direction is on the pillar to the right, which says "EXIT" with an arrow.

Common Sense Reasoning



Question: 这是哪个城市的公交车？
Which city is this bus from?

Before: 这辆公交车是广西柳州的。车身上有“柳州市公共交通集团”的标志，车牌是桂C开头，这是柳州的车牌代码。车头显示16路公交已到“犀小筑”站，车身写着“纯电动0排放”，说明是当地的新能源公交车。

This bus is from Liuzhou, Guangxi. The bus body features the "Liuzhou Public Transport Group" logo, and the license plate starts with "Gui C," which is the plate code for Liuzhou. The front of the bus shows that Route 16 has arrived at the "Xixiaozhu" station, and the body says "Pure Electric, 0 Emissions," indicating it is a local new energy bus.

After: 这辆公交车是广西桂林的16路，车身显示“桂C-03323D”，属于桂林公交集团。车身上有“纯电动0排放”的标识，说明是环保车型。从线路和车牌可以判断，这是桂林市的公交系统。