DELTA-VISION REWARD: DISTRIBUTION-AWARE MULTIMODAL REWARD MODEL

Anonymous authors
Paper under double-blind review

ABSTRACT

Multimodal reward models typically assign each response a scalar score, discarding disagreement among evaluators. We present **Delta-Vision Reward (DVR)**, which predicts a Gaussian approximation of teacher-induced reward disagreement together with a structured rationale. Distributional imitation learns the teacher ensemble's reward mean, variance, and explanatory reports; subsequent pairwise refinement targets ranking errors and insufficient preference margins. DVR achieves 72.46% accuracy on VL-Reward-Bench and 84.11% on MM-RLHF-Reward Bench. Controlled visual perturbations show that predicted disagreement responds more strongly to removal of task-relevant evidence than to equally sized irrelevant occlusions. Evaluation with an independent panel supports transfer beyond the training teachers, and distributional supervision improves ranking on two backbones under the same refinement procedure. Under a fixed revision budget, full DVR feedback achieves an 81.3% human-evaluated win rate, compared with 74.8% for rationale-only feedback. These results establish the utility of preserving teacher disagreement for multimodal reward evaluation, selective prediction, and response revision.

1 INTRODUCTION

Aligning multimodal large language models with human preferences is now a core problem in trustworthy AI (Ouyang et al., 2022; Zhang et al., 2025f; Yu et al., 2025). In current pipelines, a reward model serves as the main supervision interface between preference data and policy optimization. Its quality therefore directly affects both evaluation and alignment.

Most multimodal reward models reduce each response to a single scalar score (Zhang et al., 2025f; Zang et al., 2025; Zhang et al., 2025d; Wang et al., 2025b). This design is simple, but it discards a signal that is especially important in multimodal settings: disagreement. Preference judgments on image grounded responses are often uncertain, context dependent, and partially subjective. This is common in reasoning intensive, safety sensitive, and hallucination prone cases, where different judges may agree on the ranking but differ in confidence, or even disagree on the preferred answer altogether (Yasunaga et al., 2025; Men et al., 2025; Zhang et al., 2025g; Zhou et al., 2025a). A scalar target cannot represent this structure, and training on such collapsed supervision tends to blur hard examples into average labels.

Scalar supervision also provides little guidance on where additional preference training is needed. Process reward models introduce intermediate supervision (Wang et al., 2025a; Cao et al., 2025; Luo et al., 2025), but require finer-grained annotation. We instead retain disagreement in final-response judgments and target refinement at examples on which the student makes a ranking error or assigns an insufficient margin.

DVR predicts a reward mean, a variance, and a structured rationale for each image-query-response triple. The variance approximates disagreement among four teacher reward models; it is not an estimate of the complete human preference distribution. Distributional imitation is followed by pairwise refinement on student-teacher disagreements. The resulting student supports both mean-based ranking and variance-based abstention without requiring the teacher ensemble at inference time.

Distributional reward modeling and uncertainty-based sample selection have established precedents (Xu et al., 2025a; Sun et al., 2025). Our contribution is their integration with structured feedback in

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

multimodal reward modeling and an empirical analysis of when the learned disagreement is useful. In particular, category-level gains alone cannot establish dependence on visual evidence. We therefore intervene on the image while fixing the query and responses, and compare task-relevant and irrelevant occlusions of equal area.

Our contributions are threefold:

- An integrated multimodal reward model that retains teacher reward statistics and rationales, with refinement focused on ranking errors and low-margin pairs.
- Controlled evidence that learned disagreement responds to task-relevant visual information and transfers to evaluators outside the training ensemble.
- Evaluation of distributional supervision across backbones, teacher configurations, and feedback controls, including selective prediction and human assessment of response revision.

2 Related Work

Multimodal reward modeling. Reward modeling has become a standard component in multimodal alignment pipelines (Zhang et al., 2025f; Wang et al., 2025b; Zang et al., 2025). Recent models such as MM-RLHF-Reward, InternLM-XComposer2.5-Reward, R1-Reward, and BaseReward provide strong pairwise supervision for multimodal generation and evaluation (Zhang et al., 2025f; Zang et al., 2025; Zhang et al., 2025c;d). Related efforts also study reward models for reasoning, generation, and general purpose judging (Huang et al., 2025; Xu et al., 2025b; Xiao et al., 2025; Chen et al., 2024; Zhou et al., 2025b). Despite their differences, most of these methods learn a scalar reward. Our work departs from this setting by explicitly modeling reward uncertainty together with reward value.

Beyond scalar rewards. Distributional rewards have been studied in offline preference learning: Xu et al. (2025a) learn a reward distribution and a risk-sensitive policy using distributional Bellman updates and CVaR. Process reward models provide intermediate reasoning supervision (Wang et al., 2025a; Cao et al., 2025; Zhang et al., 2025a; Luo et al., 2025); other approaches use decomposition, adaptive feedback, or bias mitigation (Lee et al., 2025; Kim et al., 2023; Mathew & Harshit, 2025). DVR approximates the first two moments of teacher disagreement on image-conditioned responses and jointly learns structured feedback. It does not introduce a new distribution family or evaluate risk-sensitive policy optimization.

Preference optimization with informative sampling. Pairwise objectives offer an alternative to reinforcement learning based alignment (Zhang et al., 2025b), while self-improvement methods emphasize informative training examples (Thawakar et al., 2025; Hong et al., 2025; Jin et al., 2025). Closely related, Sun et al. (2025) select policy interaction samples using reward-ensemble uncertainty and gradient influence, then relabel them to refine the reward model. DVR selects offline multimodal preference pairs using student-teacher ranking errors and insufficient margins after distributional imitation. The distinction concerns the supervision and refinement setting, rather than the general principle of uncertainty-aware selection.

Benchmarks for reward evaluation. The rapid development of multimodal reward models has been accompanied by more comprehensive benchmarks, including VLRewardBench, Multimodal RewardBench, Agent-RewardBench, and domain specific variants (Li et al., 2024; Yasunaga et al., 2025; Men et al., 2025; Ding et al., 2025; Zhang et al., 2025g). These benchmarks reveal that reward modeling remains challenging in reasoning heavy, safety critical, and ambiguous scenarios. This observation further motivates our focus on disagreement-aware supervision.

3 Method

3.1 Problem Formulation

Given an image I, a query q, and a candidate response y, our reward model f_θ outputs

$$f_\theta(I, q, y) = (\mu_\theta, \sigma_\theta^2, R_\theta), \tag{1}$$

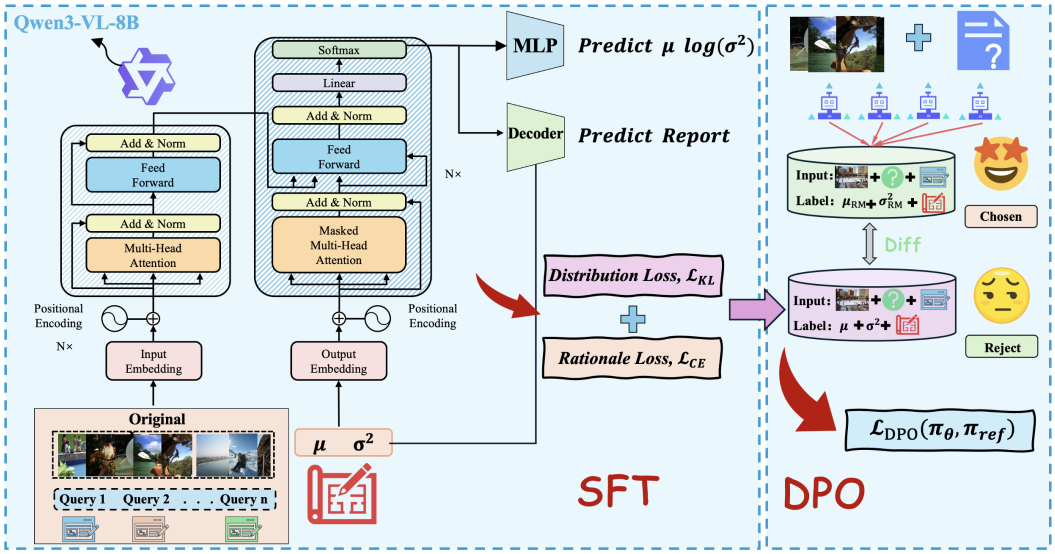

Figure 1: Overview of the DVR training framework. In the first stage, DVR learns to match a teacher induced reward distribution and to generate a structured rationale. In the second stage, we mine preference pairs on which the SFT model disagrees with the teacher consensus and refine the model with a targeted pairwise objective.

where μ_θ is the predicted reward mean, σ^2_θ is the predicted reward variance, and R_θ is a structured rationale. Unlike standard scalar reward models, DVR is trained to preserve both the central preference signal and the uncertainty of that signal. This design is motivated by the observation that many multimodal judgments are not only difficult, but also disputed.

Our method has two stages. We first train DVR to imitate a teacher induced reward distribution. We then refine its preference behavior on pairs where the student and the teacher disagree most. Figure 1 gives the full pipeline.

3.2 Teacher-Induced Distributional Supervision

We start from preference data of the form (I, q, y_w, y_l), where y_w and y_l denote the preferred and dispreferred responses. To avoid collapsing supervision into a single score, we relabel each response with a distribution estimated from a teacher ensemble.

For each response y, we collect scores $\{s_k\}_{k=1}^N$ and rationales $\{r_k\}_{k=1}^N$ from $N = 4$ multimodal reward models. We then define the teacher target as a Gaussian

$$P_T(y) = \mathcal{N}(\mu_T, \sigma_T^2), \tag{2}$$

with

$$\mu_T = \frac{1}{N}\sum_{k=1}^{N} s_k, \qquad \sigma_T^2 = \frac{1}{N-1}\sum_{k=1}^{N}(s_k - \mu_T)^2. \tag{3}$$

Here, μ_T captures the central reward signal and σ_T^2 captures teacher disagreement. With only four scores per response, the Gaussian is a parsimonious moment approximation, not an assumption that human preferences are normally distributed. A two-component mixture yields only small accuracy gains in Appendix J.4. We use the unbiased sample variance; teacher specifications and agreement statistics are given in Appendix B.

In parallel, we convert the teacher rationales into a single structured report R_T using a stronger multimodal model. This report summarizes the main evidence for the score and, when disagreement is high, explains the likely source of ambiguity. Figure 2 illustrates the labeling pipeline. The final supervision for each response is therefore the tuple (μ_T, σ_T^2, R_T).

This relabeling step serves two purposes. It yields a denser target than a binary preference label, and it makes disagreement explicit rather than treating it as annotation noise.

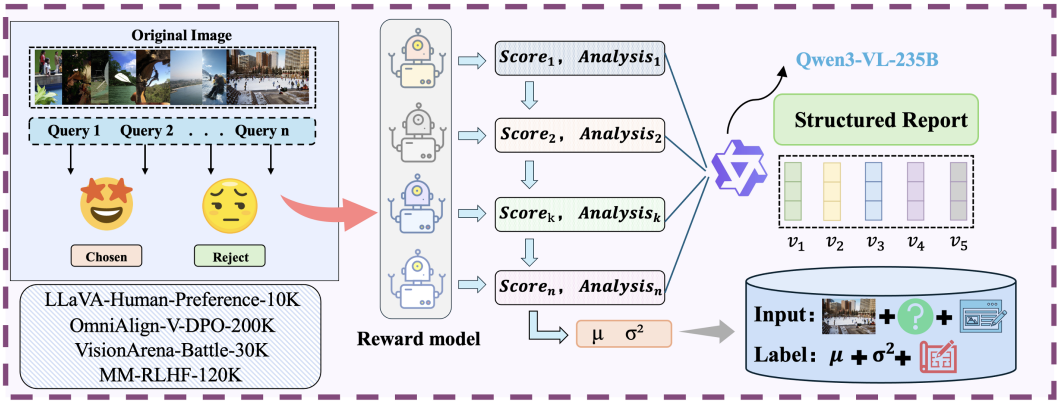

Figure 2: Distributional label construction. For each preference sample, we query a teacher ensemble to obtain multiple reward scores and rationales. The scores are aggregated into a mean and a variance, and the rationales are consolidated into a structured report. The resulting supervision preserves both preference signal and disagreement.

3.3 Stage I: Distributional Imitation

We initialize DVR from Qwen-VL-8B and attach a lightweight regression head to predict reward statistics. Given the hidden representation of (I, q, y), the head outputs μ_θ and $\log \sigma_\theta^2$, where the log-variance parameterization ensures numerical stability and nonnegative variance. The decoder generates the rationale R_θ autoregressively.

The first stage trains the model to match both the reward distribution and the explanatory report:

$$\mathcal{L}_{\text{SFT}} = \mathcal{L}_{\text{dist}} + \lambda \mathcal{L}_{\text{rat}}. \tag{4}$$

For the distribution term, we minimize the KL divergence from the teacher target to the predicted Gaussian:

$$\mathcal{L}_{\text{dist}} = D_{\text{KL}}\big(\mathcal{N}(\mu_T, \sigma_T^2) \,\|\, \mathcal{N}(\mu_\theta, \sigma_\theta^2)\big) = \log \frac{\sigma_\theta}{\sigma_T} + \frac{\sigma_T^2 + (\mu_T - \mu_\theta)^2}{2\sigma_\theta^2} - \frac{1}{2}. \tag{5}$$

This objective encourages the model to fit not only the expected reward but also the spread of judgments. In practice, this is important on hard examples where confidence and ranking quality are tightly coupled.

For the rationale term, we use standard token-level cross-entropy:

$$\mathcal{L}_{\text{rat}} = -\sum_{t=1}^{|R_T|} \log P(R_{T,t} \mid I, q, y, R_{T,<t}; \theta). \tag{6}$$

The rationale is not introduced only for interpretability. It also regularizes the reward prediction task by forcing the model to ground its judgment in explicit evidence. Additional ablations on the effect of rationale supervision are reported in Appendix E.

After this stage, the model can assign a reward mean, estimate disagreement, and provide a structured explanation. We denote the resulting model by DVR-SFT.

3.4 Stage II: Disagreement-Driven Preference Refinement

Distributional imitation gives a strong initialization, but it does not guarantee that the induced ranking behavior is optimal on difficult pairs. In particular, the student may preserve the global score scale while still making local pairwise errors. We therefore add a second stage that targets precisely those cases.

For each preference pair (y_w, y_l), we compare the student prediction to the teacher induced preference. A pair is selected if the student makes a ranking error,

$$\mu_\theta(y_w) \le \mu_\theta(y_l), \tag{7}$$

or if the student predicts the correct order with an insufficient margin,

$$\mu_\theta(y_w) - \mu_\theta(y_l) < \alpha\big(\mu_T(y_w) - \mu_T(y_l)\big), \tag{8}$$

where $\alpha \in [0, 1)$ controls how aggressively we mine low-margin pairs. The resulting subset $\mathcal{D}_{\text{hard}}$ concentrates supervision on informative failure cases rather than spending equal effort on already solved pairs. Figure 1 shows this refinement stage. A detailed comparison against alternative mining strategies is given in Appendix D.

We then continue training with a pairwise logistic objective on the mined set:

$$\mathcal{L}_{\text{pref}} = -\mathbb{E}_{(I,q,y_w,y_l)\sim\mathcal{D}_{\text{hard}}}\left[\log\sigma(\beta\left(\mu_\theta(y_w) - \mu_\theta(y_l)\right))\right], \tag{9}$$

where $\sigma(\cdot)$ is the sigmoid function and β controls the preference sharpness. This objective directly increases the score gap between preferred and dispreferred responses. Unlike uniform pairwise refinement, it focuses learning on disagreement regions where scalar reward models are typically least reliable.

The final model combines three capabilities in a single reward estimator: reward prediction, uncertainty estimation, and targeted preference discrimination.

3.5 Inference and Pairwise Decision

DVR returns $(\mu_\theta, \sigma_\theta^2, R_\theta)$ for each candidate. Public reward benchmarks use μ_θ to preserve the deterministic ranking protocol. Under an independent-Gaussian approximation, the comparison confidence is $\Phi((\mu_w - \mu_l)/\sqrt{\sigma_w^2 + \sigma_l^2})$, where Φ is the standard normal CDF. Its 0.5 decision boundary agrees with mean ranking; variance instead determines confidence and supports abstention. Appendix C evaluates this distinction.

4 Experiments

4.1 Experimental Setup

We train DVR on relabeled multimodal preference data constructed from LLaVA-Human-Preference-10K, OmniAlign-V-DPO-200K, VisionArena-Battle-30K, and MM-RLHF-120K. Evaluation is conducted on two public reward benchmarks, **VL-Reward-Bench** and **MM-RLHF-Reward Bench**, which together cover general conversation, hallucination, reasoning, safety, and video-based preference judgment. Dataset statistics, filtering rules, metric definitions, and contamination checks are reported in Appendix A and Appendix F.

We compare DVR against three groups of baselines: proprietary multimodal judges, open-source MLLMs used directly as evaluators, and specialized open-source reward models. We report pairwise preference accuracy as the primary metric. On VL-Reward-Bench, we use **Overall Acc** and **Macro Acc**. On MM-RLHF-Reward Bench, we additionally report the stricter **Acc+**, which penalizes low-margin predictions. For distribution modeling, we measure μ**MAE**, σ^2**MSE**, and **KL**. Exact definitions are given in Appendix A.

Unless otherwise stated, DVR uses Qwen-VL-8B as the backbone. The SFT and preference refinement stages are trained on 8 A100 GPUs. The main hyperparameters are $\lambda = 0.5$ and $\beta = 0.1$. We defer full training details and sensitivity studies to Appendix E.

4.2 Preference Prediction on Reward Benchmarks

We first evaluate DVR as a reward judge. The evaluation pipeline is shown in Figure 3. For each response pair, DVR predicts a reward distribution for both candidates and ranks them by the predicted mean. This setup isolates the quality of the reward model itself from downstream policy optimization.

VL-Reward-Bench. DVR achieves 72.46% overall accuracy in Table 1, compared with 71.92% for R1-Reward. It improves reasoning accuracy to 71.92%, while R1-Reward retains the highest hallucination accuracy. The overall gain therefore reflects a different balance of category performance rather than uniform dominance.

Table 1: **Preference accuracy on VL-Reward-Bench.** Best open-source reward model results are highlighted in **bold**; second-best open-source reward model results are underlined.

Model	#Param	General	Hallucination	Reasoning	Overall Acc	Macro Acc
Proprietary Models						
Claude-3.5-Sonnet	–	43.40	55.00	62.30	55.30	53.60
GPT-4o	–	49.10	67.60	70.50	65.80	62.40
Gemini-1.5-Pro	–	50.80	72.50	64.20	67.20	62.50
Claude-3.7-Sonnet	–	68.08	70.70	60.81	66.31	66.53
Open-Source MLLMs						
VITA-1.5 (Fu et al., 2025)	7B	18.55	8.93	22.11	16.48	16.53
SliME (Zhang et al., 2024)	7B	7.23	27.09	18.60	19.04	17.64
InternVL2 (Chen et al., 2023)	8B	35.60	41.10	59.00	44.50	45.20
LLaVA-Critic (Xiong et al., 2024)	8B	54.60	38.30	59.10	41.20	44.00
Molmo (Deitke et al., 2024)	72B	33.90	42.30	54.90	44.10	43.70
Qwen2-VL (Wang et al., 2024)	72B	38.10	32.80	58.00	39.50	43.00
NVLM-D (Dai et al., 2024)	72B	38.90	31.60	62.00	40.10	44.10
Llama-3.2 (Yang et al., 2024)	90B	42.60	57.30	61.70	56.20	53.90
Open-Source Reward Models						
MM-RLHF-Reward (Zhang et al., 2025e)	7B	45.04	50.45	57.55	50.15	51.01
IXC-2.5-Reward (Zang et al., 2025)	7B	<u>84.70</u>	62.50	62.90	65.80	70.00
R1-Reward (Zhao et al., 2025)	7B	63.84	**85.71**	64.78	<u>71.92</u>	<u>71.44</u>
DVR (Ours)	**8B**	65.35	<u>80.11</u>	**71.92**	**72.46**	**72.46**

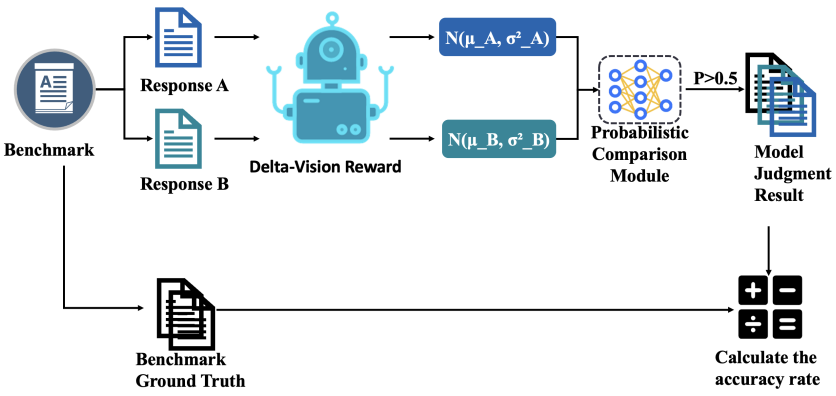

Figure 3: Preference evaluation protocol. DVR scores two candidate responses for the same image-question pair and predicts the preferred one.

MM-RLHF-Reward Bench. DVR reaches 84.11% Acc and 66.78% Acc+ in Table 2, the highest values among the compared open-source reward models. These results establish ranking performance under the common mean-based decision rule; controlled ablations below separate the effects of distributional supervision and refinement.

4.3 Distribution Alignment and Selective Prediction

We first measure agreement with the teacher-induced target distribution using the protocol in Figure 4. This tests distributional imitation, rather than calibration to an unrestricted human preference distribution.

In Table 3, DVR reduces overall KL from 1.54 for GPT-4o and 1.61 for Qwen3-VL-235B to 0.73, with corresponding reductions in mean and variance error. For selective prediction, retaining 68.1% of samples at the variance threshold $\tau = 1.0$ raises accuracy from 72.46% to 87.25% (Appendix C). This is an accuracy–coverage trade-off: the remaining 31.9% are deferred, not counted as correct. Evaluation beyond the training teachers is reported below.

Table 2: **Preference accuracy on MM-RLHF-Reward Bench.** Best open-source reward model results are highlighted in **bold**; second-best open-source reward model results are underlined.

Model	#Param	MCQ	Long	Short	Safety	Video	Acc	Acc+
Proprietary Models								
Gemini-2.0-Flash-Exp	–	33.33	45.94	67.64	43.75	32.00	44.71	13.04
GPT-4o	–	64.28	78.37	44.11	56.25	40.00	58.23	26.01
Claude-3.5-Sonnet	–	64.28	67.56	55.88	65.62	60.00	62.94	26.11
Claude-3.7-Sonnet	–	66.67	91.89	91.18	87.50	76.00	82.35	65.22
Open-Source MLLMs								
SliME (Zhang et al., 2024)	8B	23.81	10.81	14.71	12.50	7.52	17.10	1.76
VITA-1.5 (Fu et al., 2025)	7B	24.97	21.62	11.76	18.75	12.40	20.58	2.78
Intern-VL-3 (Zhu et al., 2025)	8B	35.71	56.76	23.53	37.50	32.00	37.65	6.52
NVLM-D (Dai et al., 2024)	72B	42.85	32.43	8.82	50.00	40.00	34.70	6.52
Llama-3.2 (Yang et al., 2024)	90B	19.04	35.13	38.23	50.00	40.00	35.29	10.86
Qwen2-VL (Wang et al., 2024)	72B	45.23	62.16	47.05	46.88	36.00	48.23	13.04
Open-Source Reward Models								
IXC-2.5-Reward (Zang et al., 2025)	7B	52.38	91.89	67.65	62.50	88.00	71.18	50.00
MM-RLHF-Reward (Zhang et al., 2025e)	7B	83.00	97.00	74.00	69.00	88.00	82.00	63.00
R1-Reward (Zhao et al., 2025)	7B	80.95	89.19	82.35	75.00	72.00	80.59	54.35
DVR (Ours)	**8B**	**83.52**	**97.35**	**92.10**	**88.25**	**88.50**	**84.11**	**66.78**

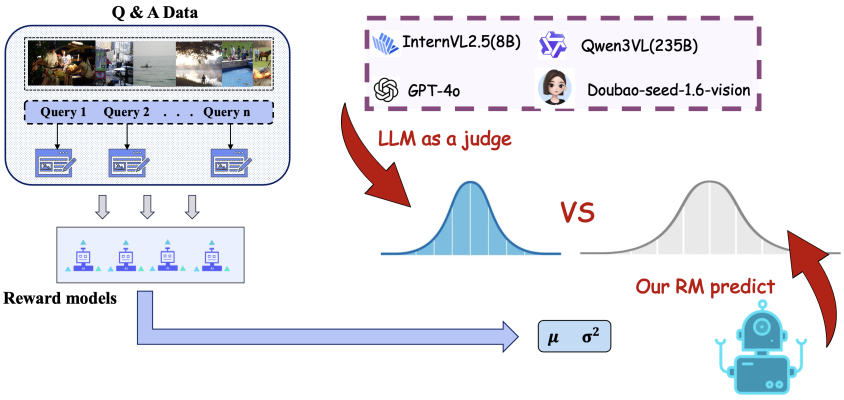

Figure 4: Distribution alignment evaluation. DVR predicts a reward distribution for each response and is compared against the teacher-induced target distribution.

4.4 Distributional Supervision and Refinement

Table 4 isolates the two training components. Distributional SFT improves VL-Reward accuracy from 68.50% to 70.20%. Refinement raises the mean-only baseline to 71.30%, while full DVR reaches 72.46%. Thus, distributional supervision adds 1.16 percentage points under the same refinement stage. On InternVL2-4B, the corresponding gain is 1.25 points (68.71% to 69.96%; Appendix J.5). Refinement accounts for the larger individual gain, but does not subsume distributional supervision. Full-model gains over mean-only SFT increase from 0.7 to 6.9 points across the lowest to highest teacher-disagreement groups; this comparison measures the combined pipeline effect.

4.5 Visual Grounding and External Validity

We perturb only the image on 312 vision-dependent samples and 84 language-dominant controls, without retraining (Table 6). On vision-dependent samples, occluding relevant evidence increases predicted variance by 0.93, versus 0.08 for equal-area irrelevant occlusion; pairwise accuracies are 64.9% and 73.7%, respectively. The corresponding variance increases on language-dominant controls are 0.05 and 0.03. This within-sample contrast supports sensitivity to task-relevant visual evidence beyond generic corruption sensitivity (Appendix J.1).

Table 3: **Distribution alignment on VL-Reward-Bench.** Lower is better. Best non-oracle results are in **bold**.

	General			Hallucination			Reasoning			Overall		
Model	μMAE	σ^2MSE	KL	μMAE	σ^2MSE	KL	μMAE	σ^2MSE	KL	μMAE	σ^2MSE	KL
Ensemble (Oracle)	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Qwen3-VL-235B	0.78	1.15	1.45	0.85	1.28	1.62	0.91	1.40	1.77	0.85	1.28	1.61
GPT-4o	0.75	1.10	1.38	0.81	1.22	1.55	0.88	1.35	1.70	0.81	1.22	1.54
InternVL2.5	0.90	1.30	1.68	0.95	1.45	1.85	1.02	1.58	2.01	0.96	1.44	1.85
Doubao-seed-1.6	0.88	1.25	1.60	0.92	1.38	1.78	0.99	1.50	1.92	0.93	1.38	1.77
DVR (Ours)	**0.48**	**0.55**	**0.61**	**0.53**	**0.68**	**0.74**	**0.59**	**0.75**	**0.83**	**0.53**	**0.66**	**0.73**

Table 4: **Ablation on VL-Reward-Bench.** Each component contributes consistently, and the full model performs best.

Configuration	Dist.	Refine	Overall Acc (%)
DVR (μ-only, SFT-only)	✗	✗	68.50
DVR (SFT-only)	✓	✗	70.20
DVR (μ-only, + pairwise refine)	✗	✓	71.30
DVR (Full)	✓	✓	**72.46**

With an evaluator panel excluded from label construction, DVR achieves 79.6% pairwise agreement and 70.4% agreement on hard cases, versus 74.8% and 61.5% for a mean-only reward model. These results provide external evidence for reward quality, while leaving calibration to the full distribution of human disagreement unresolved (Appendix J.3).

4.6 DVR AS A COACH

DVR critiques an initial LLaVA-1.5-7B response and guides one revision (Figure 5). Table 5 reports the original comparison with unrevised responses. Since that comparison also changes the inference budget and uses judges overlapping with the teacher ensemble, it cannot isolate the contribution of feedback.

We therefore compare feedback types using the same base model, revision prompt, and one revision, with budgets of 128 critique tokens and 256 revision tokens. Full DVR feedback achieves an 84.1% win rate, versus 56.9% for self-revision, 69.4% for scalar feedback, and 77.7% for rationale-only feedback. Under human evaluation, the respective rates are 81.3%, 54.0%, 66.5%, and 74.8%. Average revision lengths are similar across feedback conditions (242–245 tokens). Appendix J.8 reports the controls and independent-judge results.

5 DISCUSSION

Scope of the contribution. DVR combines established distributional modeling and targeted refinement with multimodal supervision and structured feedback. Visual interventions support dependence on task-relevant evidence, while independent evaluation and controlled coaching support utility beyond teacher imitation. These findings do not establish a new uncertainty estimator or eliminate shared teacher bias.

Table 5: **Coach evaluation.** LLaVA-1.5-7B revises its response using DVR feedback. Win rate is measured by the original reward-model panel.

Evaluator	General (%)	Hallucination (%)	Reasoning (%)	Overall Win Rate (%)
MM-RLHF-Reward (Zhang et al., 2025e)	78.5	81.2	82.0	80.57
IXC-2.5-Reward (Zang et al., 2025)	91.2	88.5	89.4	89.70
R1-Reward (Zhao et al., 2025)	82.1	88.9	82.0	84.32
Average	**82.9**	**85.0**	**84.2**	**84.03**

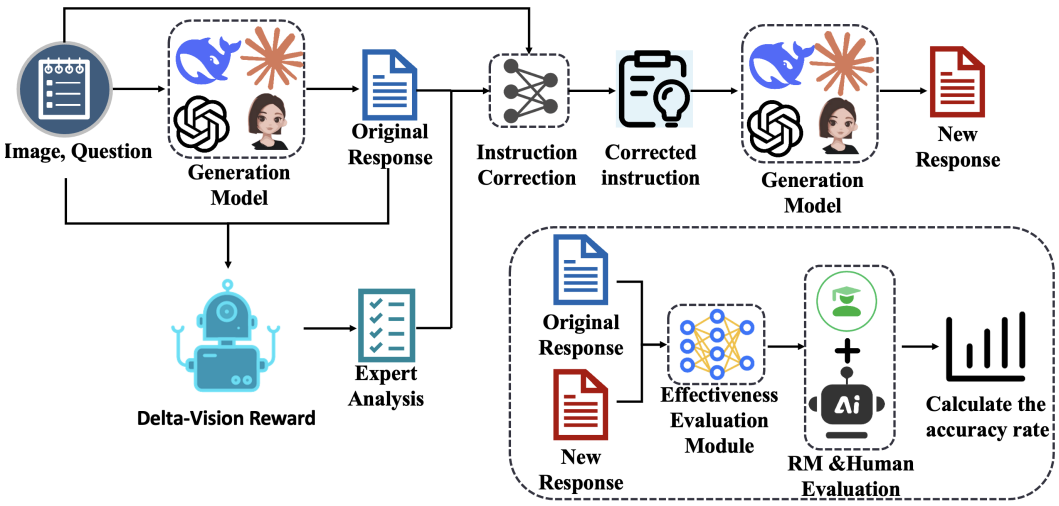

Figure 5: Coach evaluation. DVR critiques an initial response, and the critique is used to guide revision by a base MLLM.

Table 6: Effect of equal-area occlusion on predicted variance and pairwise accuracy. Variance changes are relative to the clean image.

Sample group	Occluded region	$\Delta\sigma_\theta^2$	Acc. (%)
Vision-dependent ($n = 312$)	Irrelevant	+0.08	73.7
	Task-relevant	+0.93	64.9
Language-dominant ($n = 84$)	Irrelevant	+0.03	78.1
	Task-relevant	+0.05	77.9

Cost and limitations. Teacher labeling is an offline cost; deployment uses one student. The measured full labeling run takes 36.8 GPU-hours and 5.2 wall-clock hours, at 6.13 times the scalar baseline cost. A three-teacher variant without rationale merging reaches 71.84% VL-Reward accuracy at 3.70 times baseline cost (Appendix J.7). The Gaussian approximation cannot represent multimodal disagreement, and external agreement does not identify the full human preference distribution. Selective prediction trades coverage for accuracy; downstream policy optimization and end-to-end human escalation are not evaluated. Contamination checks and failure cases are documented in Appendices F and H.

6 Conclusion

DVR preserves teacher-induced reward disagreement and structured rationales, then refines ranking on difficult preference pairs. The evaluations support improved ranking across two backbones, sensitivity to visual evidence, transfer to independent evaluators, and feedback utility under a controlled revision budget. The model provides a practical basis for selective multimodal evaluation, subject to the limitations of its teachers and distributional approximation.

AI USE STATEMENT

The method described in this paper uses multimodal generative models to construct teacher-derived reward statistics and structured rationales, and to generate and revise responses in the coach evaluation. The corresponding protocols are documented in Sections 3 and 4.6 and Appendices B and I. A generative AI assistant was also used to migrate the manuscript to the ICLR LaTeX template and assist with formatting and consistency checks.

REPRODUCIBILITY STATEMENT

Benchmark specifications, metric definitions, and evaluation procedures are provided in Appendix A. Teacher construction, uncertainty validation, sample mining, and ablation settings are documented in Appendices B–E. Contamination checks, qualitative cases, limitations, and the coach evaluation protocol are provided in Appendices F–I. Controlled visual perturbations, external evaluation, teacher sensitivity, and matched-budget feedback comparisons are documented in Appendix J.

REFERENCES

Qi Cao, Ruiyi Wang, Ruiyi Zhang, Sai Ashish Somayajula, and Pengtao Xie. Dreamprm: Domain-reweighted process reward model for multimodal reasoning. *arXiv preprint arXiv:2505.20241*, 2025.

Zhaorun Chen, Yichao Du, Zichen Wen, Yiyang Zhou, Chenhang Cui, Zhenzhen Weng, Haoqin Tu, Chaoqi Wang, Zhengwei Tong, Qinglan Huang, et al. Mj-bench: Is your multimodal reward model really a good judge for text-to-image generation? *arXiv preprint arXiv:2407.04842*, 2024.

Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, Bin Li, Ping Luo, Tong Lu, Yu Qiao, and Jifeng Dai. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. arXiv, 2023. Corresponds to InternVL2 (Chen et al., 2023b) in the table.

Wenliang Dai, Nayeon Lee, Boxin Wang, Zhuolin Yang, Zihan Liu, Jon Barker, Tuomas Rintamaki, Mohammad Shoeybi, Bryan Catanzaro, and Wei Ping. Nvlm: Open frontier-class multimodal llms. arXiv, 2024. Corresponds to NVLM-D (Dai et al., 2024) in the table.

Matt Deitke, Christopher Clark, Sangho Lee, Rohun Tripathi, Yue Yang, Jae Sung Park, Mohammadreza Salehi, Niklas Muennighoff, Kyle Lo, Luca Soldaini, et al. Molmo and pixmo: Open weights and open data for state-of-the-art multimodal models. arXiv, 2024. Corresponds to Molmo (Deitke et al., 2024) in the table.

Meidan Ding, Jipeng Zhang, Wenxuan Wang, Cheng-Yi Li, Wei-Chieh Fang, Hsin-Yu Wu, Haiqin Zhong, Wenting Chen, and Linlin Shen. Med-rewardbench: Benchmarking reward models and judges for medical multimodal large language models. *arXiv preprint arXiv:2508.21430*, 2025.

Chaoyou Fu, Haojia Lin, Xiong Wang, Yi-Fan Zhang, Yunhang Shen, Xiaoyu Liu, Yangze Li, Zuwei Long, Heting Gao, Ke Li, et al. Vita-1.5: Towards gpt-4o level real-time vision and speech interaction. arXiv, 2025. Corresponds to VITA-1.5 (Fu et al., 2025) in the table.

Jixiang Hong, Yiran Zhang, Guanzhong Wang, Yi Liu, Ji-Rong Wen, and Rui Yan. Suder: Self-improving unified large multimodal models for understanding and generation with dual self-rewards, 2025. URL https://arxiv.org/abs/2506.07963.

Wenxuan Huang, Bohan Jia, Zijie Zhai, Shaosheng Cao, Zheyu Ye, Fei Zhao, Zhe Xu, Yao Hu, and Shaohui Lin. Vision-r1: Incentivizing reasoning capability in multimodal large language models. *arXiv preprint arXiv:2503.06749*, 2025.

Weiyang Jin, Yuwei Niu, Jiaqi Liao, Chengqi Duan, Aoxue Li, Shenghua Gao, and Xihui Liu. Srum: Fine-grained self-rewarding for unified multimodal models. *arXiv preprint arXiv:2510.12784*, 2025.

Changyeon Kim, Younggyo Seo, Hao Liu, Lisa Lee, Jinwoo Shin, Honglak Lee, and Kimin Lee. Guide your agent with adaptive multimodal rewards. *Advances in Neural Information Processing Systems*, 36:54472–54495, 2023.

Dong Won Lee, Hae Won Park, Cynthia Breazeal, and Louis-Philippe Morency. Aligning dialogue agents with global feedback via large language model reward decomposition. *arXiv preprint arXiv:2505.15922*, 2025.

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. Vlrewardbench: A challenging benchmark for vision-language generative reward models. arXiv, 2024.

Ruilin Luo, Zhuofan Zheng, Lei Wang, Yifan Wang, Xinzhe Ni, Zicheng Lin, Songtao Jiang, Yiyao Yu, Chufan Shi, Ruihang Chu, et al. Unlocking multimodal mathematical reasoning via process reward model. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

Sheryl Mathew and N Harshit. Counterfactual reward model training for bias mitigation in multimodal reinforcement learning. *arXiv preprint arXiv:2508.19567*, 2025.

Tianyi Men, Zhuoran Jin, Pengfei Cao, Yubo Chen, Kang Liu, and Jun Zhao. Agent-rewardbench: Towards a unified benchmark for reward modeling across perception, planning, and safety in real-world multimodal agents. *arXiv preprint arXiv:2506.21252*, 2025.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*, volume 35, 2022.

Zexu Sun, Yiju Guo, Yankai Lin, Xu Chen, Qi Qi, Xing Tang, Xiuqiang He, and Ji-Rong Wen. Uncertainty and influence aware reward model refinement for reinforcement learning from human feedback. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/fd7259e22add6de6df8ff0ccc902a34d-Abstract-Conference.html.

Omkat Thawakar, Shravan Venkatraman, Ritesh Thawkar, Abdelrahman Shaker, Hisham Cholakkal, Rao Muhammad Anwer, Salman Khan, and Fahad Khan. Evolmm: Self-evolving large multimodal models with continuous rewards. *arXiv preprint arXiv:2511.16672*, 2025.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. arXiv, 2024. Corresponds to Qwen2-VL (Wang et al., 2024b) in the table.

Weiyun Wang, Zhangwei Gao, Lianjie Chen, Zhe Chen, Jinguo Zhu, Xiangyu Zhao, Yangzhou Liu, Yue Cao, Shenglong Ye, Xizhou Zhu, et al. Visualprm: An effective process reward model for multimodal reasoning. *arXiv preprint arXiv:2503.10291*, 2025a.

Xiaokun Wang, Peiyu Wang, Jiangbo Pei, Wei Shen, Yi Peng, Yunzhuo Hao, Weijie Qiu, Ai Jian, Tianyidan Xie, Xuchen Song, et al. Skywork-vl reward: An effective reward model for multimodal understanding and reasoning. *arXiv preprint arXiv:2505.07263*, 2025b.

Tong Xiao, Xin Xu, Zhenya Huang, Hongyu Gao, Quan Liu, Qi Liu, and Enhong Chen. Advancing multimodal reasoning capabilities of multimodal large language models via visual perception reward. *arXiv preprint arXiv:2506.07218*, 2025.

Tianyi Xiong, Xiyao Wang, Dong Guo, Qinghao Ye, Haoqi Fan, Quanquan Gu, Heng Huang, and Chunyuan Li. Llava-critic: Learning to evaluate multimodal models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. Corresponds to LLaVA-Critic (Xiong et al., 2024) in the table.

Sheng Xu, Bo Yue, Hongyuan Zha, and Guiliang Liu. A distributional approach to uncertainty-aware preference alignment using offline demonstrations. In *The Thirteenth International Conference on Learning Representations*, 2025a. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/084cf2b3d73abafa1705336a0e9ebb1c-Abstract-Conference.html.

Shilin Xu, Yanwei Li, Rui Yang, Tao Zhang, Yueyi Sun, Wei Chow, Linfeng Li, Hang Song, Qi Xu, Yunhai Tong, et al. Mixed-r1: Unified reward perspective for reasoning capability in multimodal large language models. *arXiv preprint arXiv:2505.24164*, 2025b.

Minghao Yang, Chao Qu, and Xiaoyu Tan. Inf-orm-llama3.1-70b. https://huggingface.co/infly/INF-ORM-Llama3.1-70B, 2024. Corresponds to Llama-3.2 (Minghao Yang, 2024) in the table.

Michihiro Yasunaga, Luke Zettlemoyer, and Marjan Ghazvininejad. Multimodal rewardbench: Holistic evaluation of reward models for vision language models. *arXiv preprint arXiv:2502.14191*, 2025.

Tao Yu, Chaoyou Fu, Junkang Wu, Jinda Lu, Kun Wang, Xingyu Lu, Yunhang Shen, Guibin Zhang, Dingjie Song, Yibo Yan, et al. Aligning multimodal llm with human preference: A survey. arXiv, 2025.

Yuhang Zang, Xiaoyi Dong, Pan Zhang, Yuhang Cao, Ziyu Liu, Shengyuan Ding, Shenxi Wu, Yubo Ma, Haodong Duan, Wenwei Zhang, et al. Internlm-xcomposer2. 5-reward: A simple yet effective multi-modal reward model. *arXiv preprint arXiv:2501.12368*, 2025.

Jianghangfan Zhang, Yibo Yan, Kening Zheng, Xin Zou, Song Dai, and Xuming Hu. Gm-prm: A generative multimodal process reward model for multimodal mathematical reasoning. *arXiv preprint arXiv:2508.04088*, 2025a.

Jiawei Zhang, Tianyu Pang, Chao Du, Yi Ren, Bo Li, and Min Lin. Benchmarking large multimodal models against common corruptions. arXiv, 2024. Corresponds to SliME (Zhang et al., 2024b) in the table.

Ruohong Zhang, Liangke Gui, Zhiqing Sun, Yihao Feng, Keyang Xu, Yuanhan Zhang, Di Fu, Chunyuan Li, Alexander G Hauptmann, Yonatan Bisk, et al. Direct preference optimization of video large multimodal models from language model reward. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 694–717, 2025b.

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

Yi-Fan Zhang, Haihua Yang, Huanyu Zhang, Yang Shi, Zezhou Chen, Haochen Tian, Chaoyou Fu, Haotian Wang, Kai Wu, Bo Cui, et al. Basereward: A strong baseline for multimodal reward model. *arXiv preprint arXiv:2509.16127*, 2025d.

Yi-Fan Zhang, Tao Yu, Haochen Tian, Chaoyou Fu, Peiyan Li, Jianshu Zeng, Wulin Xie, Yang Shi, Huanyu Zhang, Junkang Wu, et al. Mm-rlhf: The next step forward in multimodal llm alignment. arXiv, 2025e. Corresponds to MM-RLHF-Reward (Zhang et al., 2025a) in the table.

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

Zhihong Zhang, Xiaojian Huang, Jin Xu, Zhuodong Luo, Xinzhi Wang, Jiansheng Wei, and Xuejin Chen. Videorewardbench: Comprehensive evaluation of multimodal reward models for video understanding. *arXiv preprint arXiv:2509.00484*, 2025g.

Jiaxing Zhao, Xihan Wei, and Liefeng Bo. R1-omni: Explainable omni-multimodal emotion recognition with reinforcing learning. arXiv, 2025. Corresponds to R1-Reward (Zhao et al., 2025b) in the table.

Pengfei Zhou, Jiaxin Ai, Wangbo Zhao, Kai Wang, Xiaojiang Peng, Wenqi Shao, Hongxun Yao, Kaipeng Zhang, et al. Mpbench: A comprehensive multimodal reasoning benchmark for process errors identification. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 21586–21606, 2025a.

Shijie Zhou, Ruiyi Zhang, Huaisheng Zhu, Branislav Kveton, Yufan Zhou, Jiuxiang Gu, Jian Chen, and Changyou Chen. Multimodal llms as customized reward models for text-to-image generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 19638–19648, 2025b.

Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Yuchen Duan, Hao Tian, Weijie Su, Jie Shao, et al. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. arXiv, 2025. Corresponds to Intern-VL-3 (Zhu et al., 2025) in the table.