RANKING-AWARE CALIBRATION FOR RELIABLE MULTIMODAL REINFORCEMENT LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Reinforcement learning (RL) post-training has substantially improved the reasoning accuracy of vision-language models, yet the resulting policies remain poorly calibrated. Terminal correctness rewards provide no gradient that penalizes confident errors more than uncertain ones, and no signal that ties confidence to the quality of the supporting visual evidence. This mismatch becomes especially severe under corrupted or ambiguous visual inputs, where models continue to report high confidence on wrong answers. We introduce **R**anking-**A**ware **C**alibration (RAC), a training-time framework that supervises confidence using two comparison signals that group-based RL already produces at no additional labeling cost. The *ranking-aware group loss* enforces that a better rollout receives higher confidence than a worse one within the same prompt. The *clean–corrupted pairwise loss* enforces that confidence attenuates as visual evidence degrades. The ranking signal requires the policy to distinguish between correct and incorrect rollouts and improves task accuracy on most evaluated backbones. Both losses require no external confidence annotations and integrate naturally with group-based RL post-training. We instantiate RAC on Qwen2.5-VL, Qwen3.5, and InternVL-3.5 backbones and evaluate on six multimodal reasoning benchmarks under clean and corrupted inputs. Empirical results show that the ranking-aware loss improves task accuracy on most evaluated backbones, while the pairwise corruption loss reduces calibration error. Their combination improves ECE and Brier score over Vanilla-RL on all four backbones and achieves best or joint-best calibration among the compared methods across all four, while improving accuracy on three of the four backbones. Our code is available at `https://anonymous.4open.science/r/RAC`.

1 INTRODUCTION

The integration of visual perception and language reasoning has driven rapid progress in artificial intelligence. Modern vision-language models (VLMs) and multimodal large language models (MLLMs) can now solve tasks ranging from spatial understanding to multi-step mathematical reasoning (Radford et al., 2021; Alayrac et al., 2022; Li et al., 2023a; Liu et al., 2023; Bai et al., 2023; Wang et al., 2024b; Bai et al., 2025; Zhu et al., 2025). A major driver of these gains has been reinforcement learning (RL) with verifiable outcome rewards: by optimizing directly for correctness, models learn to produce longer reasoning traces, revise intermediate conclusions, and self-correct, yielding substantial performance improvements on visual mathematics and science benchmarks (Team, 2024; AI, 2025; Wang et al., 2025b).

However, better reasoning accuracy does not automatically translate into better calibration. This problem has since been studied extensively in text-only LLMs through several distinct directions, including post-hoc recalibration of model probabilities via temperature scaling and adaptive variants (Kadavath et al., 2022; Tian et al., 2023; Xie et al., 2024), verbalized confidence elicitation (Xiong et al., 2024; Zhang et al., 2024), and training-time interventions such as representation editing (Ji et al., 2025), listener-aware fine-tuning (Stengel-Eskin et al., 2024), calibration-aware reward shaping during RL post-training (Damani et al., 2025), and analyses of how chain-of-thought behavior itself affects calibration (Yoon et al., 2025). By contrast, calibration in multimodal reasoning models remains comparatively underexplored, and recent evidence suggests that confidence can become especially fragile when visual evidence is degraded or ambiguous (Saxena et al., 2026; Sui et al., 2025). Prior work has shown that calibration often degrades under distribution shift and input corruption (Ova-

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.

dia et al., 2019; Hendrycks & Dietterich, 2019; Minderer et al., 2021). An MLLM may inspect a corrupted image, generate a fluent rationale, arrive at an incorrect answer, and still report high confidence. The underlying issue is structural: terminal correctness rewards distinguish right answers from wrong ones, but they do not explicitly penalize *confident errors* more than uncertain ones, nor do they require confidence to track the quality of the supporting visual evidence.

This gap is especially damaging because it directly amplifies hallucination. When a model reasons confidently from weak or corrupted visual input, the failure is not merely an isolated prediction error but a form of authoritative misreasoning that can propagate to downstream decision-making systems (Li et al., 2023b; Chen et al., 2024b; Dang et al., 2026). Recent analyses show that VLMs frequently assign high verbalized confidence to hallucinated answers, and that the gap between stated certainty and actual correctness widens sharply under visual corruption (Dang et al., 2026; Byun et al., 2026). In systems that use confidence for abstention, reranking, or query routing, this failure mode is particularly costly. Existing methods fail to address the core challenge in multimodal reasoning: confidence should respond to the quality of visual evidence. It should fall when evidence quality falls, and among multiple candidate solutions to the same question, better reasoning should receive higher confidence than worse reasoning. These directional properties are absent from text-only objectives and cannot be retroactively imposed by post-hoc rescaling of a frozen model. Moreover, learning to rank confidence across rollouts of the same prompt provides an additional discrimination signal, with observed accuracy gains over correctness-only rewards on most evaluated backbones.

We therefore treat calibration as a native objective within multimodal RL. Our key observation is that group-based RL already provides two forms of supervision at no additional labeling cost. For each prompt, the policy samples multiple rollouts with known correctness, yielding a *within-prompt* comparison axis. Across prompts, any clean input can be paired with a synthetically corrupted version of the same question, yielding an *across-evidence* comparison axis. These two axes capture exactly the directional properties missing from standard outcome rewards, without requiring any external confidence annotations. We instantiate this idea in the **RAC** (**R**anking-**A**ware **C**alibration) framework, which encodes the two axes as margin-based losses integrated directly into the RL reward. The *ranking-aware group loss* enforces that, within the same prompt, a better rollout receives higher confidence than a worse one. The *clean—corrupted pairwise loss* enforces that confidence decreases when the same question is presented with weaker visual evidence, with the required gap scaling with corruption severity. Together, these objectives teach the model to align confidence with both reasoning quality and evidence quality during RL training itself, and can also improve task accuracy, as observed on most evaluated backbones.

Our contributions are threefold:

- **A relative supervision view of RL calibration.** We identify two directional properties that standard outcome-based multimodal RL fails to enforce, and show that both can be supervised from comparisons already produced by RL sampling, namely within-prompt rollout groups and matched clean–corrupted pairs, without requiring external confidence labels.
- **The RAC framework.** We formalize these two supervision axes as margin-based ranking and pairwise losses and distribute them back into sequence-level rewards, yielding a unified calibration objective that integrates directly into any group-based RL post-training pipeline without requiring changes to the underlying optimizer or model architecture.
- **Empirical evaluation under visual corruption.** We instantiate RAC on Qwen2.5-VL-3B/7B-Instruct, Qwen3.5-4B, and InternVL-3.5-8B, and evaluate it on six multimodal reasoning benchmarks under clean and progressively corrupted inputs, covering eight ImageNet-C operators and five severity levels. The ranking-aware loss improves accuracy on most backbones. The combined objectives improve ECE and Brier score over Vanilla-RL on all four backbones, with accuracy gains on three of them.

2 Related Work

Outcome-based RL supports multimodal reasoning (Team, 2024; Wang et al., 2025b). Calibration methods include post-hoc rescaling (Guo et al., 2017), verbalized confidence elicitation (Xiong et al., 2024), and training-time supervision (Stengel-Eskin et al., 2024; Damani et al., 2025). Visual corruption further challenges calibration (Ovadia et al., 2019). RAC combines within-prompt

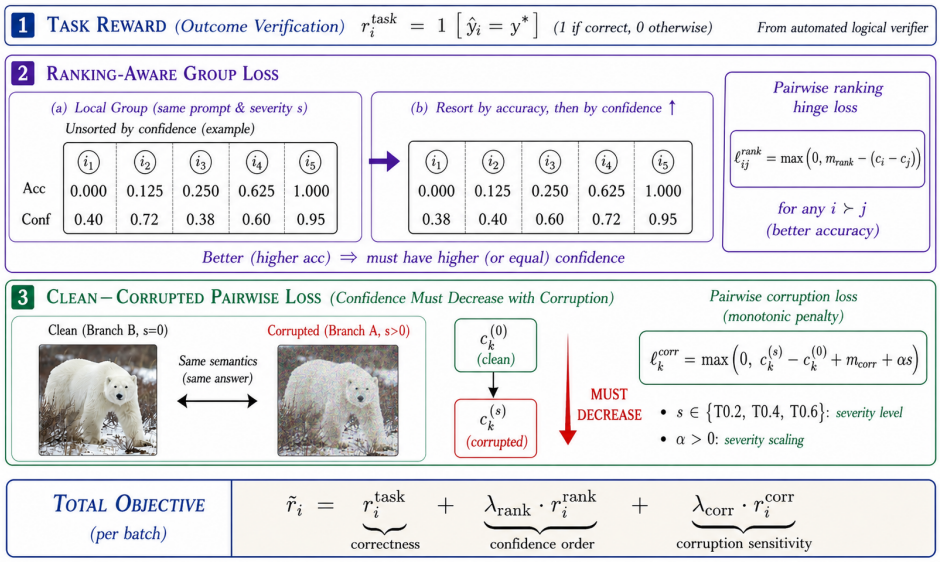


Figure 1: **Overview of the RAC Calibration Architecture.** The framework is structured around three reinforcement signals: (1) the standard outcome verification task reward, (2) the Ranking-Aware Group Loss, which asserts that higher confidence should coordinate with greater answer correctness within the same rollout group, and (3) the Clean–Corrupted Pairwise Loss, which enforces directional confidence attenuation when the visual input is synthetically degraded.

correctness comparisons with clean–corrupted pairs to supervise verbalized confidence during RL. Appendix A provides additional related work.

3 METHOD

This section presents the RAC framework in three parts. We first review the outcome-based multimodal RL setting and the group-based sampling geometry on which our method relies (Section 3.1). We then introduce the two RAC calibration objectives: one for ordering confidence within a prompt, and one for enforcing confidence attenuation across clean and corrupted views (Section 3.2). Finally, we explain how these objectives are converted into shaped rewards and optimized in a unified training loop (Section 3.3).

3.1 OUTCOME-BASED REINFORCEMENT LEARNING

In outcome-based multimodal RL, the goal is to fine-tune a pre-trained policy π_θ to maximize verifier-derived rewards. Classical alignment methods often relied on Proximal Policy Optimization (PPO), which stabilizes policy updates using a learned value baseline V_ϕ. However, training a full multimodal critic that mirrors the policy can introduce substantial memory overhead for modern VLMs.

Recent sample-efficient RL methods avoid the explicit critic by using multiple rollouts from the same prompt to construct relative baselines (Team, 2024; Ahmadian et al., 2024). This paradigm has since been extended to multimodal reasoning, with recent work adapting group-based RL to train vision-language models on verifiable visual tasks (Wang et al., 2025a; Huang et al., 2025). Let $x \in \mathcal{D}$ denote a multimodal prompt containing text and image input. For each prompt, the policy samples a group of n rollouts $\mathcal{G}(x) = \{y_1, \ldots, y_n\}$ and receives task rewards $r_1, \ldots, r_n$ from a verifier. This prompt-level rollout group is the basic structural unit on which RAC operates.

Group Relative Policy Optimization (GRPO) (Team, 2024) directly estimates the baseline by standardizing the returns within the group $\mathcal{G}(x)$. The advantage for a rollout y_i is computed as:

$$A_i^{\text{GRPO}} = \frac{r_i - \mu_r}{\sigma_r}, \qquad \mu_r = \frac{1}{n}\sum_{j=1}^{n} r_j, \qquad \sigma_r = \sqrt{\frac{1}{n}\sum_{j=1}^{n}(r_j - \mu_r)^2}. \tag{1}$$

GRPO supports a standard clipped policy optimization objective with a reference model regularizer:

$$\mathcal{J}_{\text{RL}}(\theta) = \mathbb{E}_{x\sim\mathcal{D},\ \{y_i\}_{i=1}^{n}\sim\pi_{\text{old}}}\left[\frac{1}{n}\sum_{i=1}^{n}\min\big(\rho_i(\theta)A_i,\ \text{clip}\big(\rho_i(\theta), 1-\epsilon, 1+\epsilon\big)A_i\big)\right] - \beta_{\text{KL}}\,\mathbb{D}_{\text{KL}}(\pi_\theta\|\pi_{\text{ref}}), \tag{2}$$

where $\rho_i(\theta) = \frac{\pi_\theta(y_i|x)}{\pi_{\text{old}}(y_i|x)}$ is the importance ratio, ϵ controls the clipping range, and β_{KL} weights deviation from the reference model. Because GRPO already organizes optimization around prompt-level sets of alternative trajectories, it provides exactly the comparison structure needed to supervise confidence *relatively* rather than absolutely. This structure is precisely what RAC exploits: rollout groups provide the within-prompt comparisons used by the ranking loss, while clean–corrupted pairs provide the across-evidence comparisons used by the corruption loss.

3.2 Ranking-aware Calibration Objectives

Confidence-bearing rollouts. RAC assumes that each sampled rollout contains not only a rationale and final answer, but also an explicit scalar confidence value. Here, confidence denotes answer-level verbalized confidence parsed from the `<confidence>` tag. Concretely, for a prompt x, one rollout sampled from π_θ produces a completion y_i, a parsed confidence $c_i \in [0,1]$, a verifier-based correctness signal $a_i \in \{0,1\}$, and a format indicator $f_i \in \{0,1\}$ that checks whether the response follows the required output schema. We therefore represent one rollout as (y_i, c_i, a_i, f_i) and sample a prompt-level group

$$\mathcal{G}(x) = \{(y_i, c_i, a_i, f_i)\}_{i=1}^{n}, \qquad (y_i, c_i) \sim \pi_\theta(\cdot \mid x). \tag{3}$$

Within a single prompt, question difficulty is largely controlled. As a result, comparing rollouts from the same question provides a clean local signal for confidence quality: if several candidate solutions compete for one prompt, the better solution should be assigned higher confidence than the worse one. This relative supervision is more trustworthy than forcing each confidence value toward an absolute probability target, since no calibrated ground-truth target is available in outcome-reward RL.

Ranking-aware group calibration. For each rollout group $\mathcal{G}(x)$ (Eq. 3), we define the valid comparison set

$$\mathcal{Q}(x) = \{(i,j) : a_i > a_j,\ f_i = 1,\ f_j = 1\}. \tag{4}$$

Each pair compares a better rollout against a worse rollout from the same prompt. We then impose a margin-based ranking loss

$$\ell_{ij}^{\text{rank}} = \max\big(0,\ m_{\text{rank}} - (c_i - c_j)\big), \tag{5}$$

where $m_{\text{rank}} > 0$ is a confidence margin. The loss vanishes only when the better rollout receives sufficiently higher confidence. Aggregating over all valid pairs (Eq. 5) yields

$$L_{\text{rank}}(x) = \frac{1}{|\mathcal{Q}(x)|}\sum_{(i,j)\in\mathcal{Q}(x)} \ell_{ij}^{\text{rank}}, \tag{6}$$

with the convention that $L_{\text{rank}}(x) = 0$ when no valid comparison exists. In practice, this groupwise loss is converted into rollout-level shaping terms by accumulating, for each sampled rollout, the ranking violations of all valid pairs in which it participates.

Remark 1 (Why relative, not absolute?)**.** *Forcing confidence toward absolute targets (e.g., $c_i \to a_i$) is appealing but suffers from a mismatch: the binary correctness label a_i is not itself a well-calibrated probability target, especially for borderline or partially correct responses. In contrast, the pairwise margin only requires that better reasoning attract more certainty than worse reasoning* within the same context*. This relative objective is both weaker and better aligned with the information available from online RL rollouts. Furthermore, enforcing relative ordering within a prompt group provides a stronger training signal than a correctness reward alone: the policy must not only produce a correct answer but also assign it distinguishably higher confidence than incorrect alternatives, which directly reinforces the quality of the underlying reasoning.*

Clean–corrupted pairwise calibration. The ranking term governs confidence *within* a prompt, but it does not say how confidence should change when the quality of the underlying evidence changes. We therefore introduce paired clean and corrupted views $x^{(0)}$ and $x^{(s)}$ of the same question, where $s \in [0, 1]$ denotes corruption severity and $s = 0$ is the clean input. The desired behavior is directional: if two inputs share the same semantic content but one contains weaker visual evidence, the corrupted view should not retain the same level of confidence as the clean one.

For each clean prompt $x^{(0)}$ and its corrupted counterpart $x^{(s)}$, we sample two rollout groups. GRPO baselines and advantage normalization are computed separately within each group; cross-view pairing is used only for auxiliary reward shaping. We then form slot-wise pairs

$$\big((y_k^{(0)}, c_k^{(0)}, a_k^{(0)}), (y_k^{(s)}, c_k^{(s)}, a_k^{(s)})\big). \tag{7}$$

Here, the index k identifies corresponding draws from the two groups inside the same training iteration. It is used to organize the pairwise supervision and does not assume token-level trajectory correspondence between clean and corrupted rollouts. We then impose a directional margin that favors higher confidence on the clean view:

$$\ell_k^{\text{corr}} = \max\big(0,\ c_k^{(s)} - c_k^{(0)} + m_{\text{corr}} + \alpha s\big), \tag{8}$$

where $m_{\text{corr}} > 0$ is a base margin and $\alpha \geq 0$ scales the desired confidence drop with corruption severity. This objective does not require every corrupted sample to be low-confidence in isolation. Rather, it expresses a cleaner comparative rule: holding semantic content fixed, the corrupted view should not be assigned the same or higher confidence than the clean view without incurring a penalty.

Aggregating over the slot-wise pairs (Eq. 8) yields:

$$L_{\text{corr}}(x^{(0)}, x^{(s)}) = \frac{1}{n} \sum_{k=1}^{n} \ell_k^{\text{corr}}. \tag{9}$$

Uniformly lowering all confidence values for a corrupted prompt may reduce the pairwise corruption loss, but leaves its within-prompt confidence differences $c_i - c_j$ unchanged. It therefore cannot, by itself, reduce within-prompt ranking violations.

Connection to contrastive learning. The corruption loss is structurally similar in spirit to supervised contrastive objectives (Khosla et al., 2020), which use relative relationships rather than absolute labels. Here, however, the comparison is imposed directly on the scalar confidence rather than on a representation vector: the clean and corrupted views of the same question form a matched pair, and the clean view is required to be *more confident*. This makes it possible to supervise calibration without any explicit confidence annotation.

3.3 Optimization and Training Procedure

Unified objective. The two losses play complementary roles. The ranking term specifies how confidence should be ordered among alternative solutions to the same prompt, while the corruption term specifies how confidence should change when evidence quality degrades. Their combined calibration objective is

$$\mathcal{L}_{\text{cal}} = \mathbb{E}_{x \sim \mathcal{D}}\big[L_{\text{rank}}(x)\big] + \beta\, \mathbb{E}_{(x^{(0)}, x^{(s)}) \sim \mathcal{D}_{\text{pair}}}\big[L_{\text{corr}}(x^{(0)}, x^{(s)})\big], \tag{10}$$

where $\mathcal{D}$ is the prompt distribution, $\mathcal{D}_{\text{pair}}$ is the paired clean–corrupted distribution, and β balances the two relative supervision signals. In practice, we optimize a sequence-level RL objective whose shaped reward is

$$\tilde{r}_i = r_i^{\text{task}} + \lambda_{\text{rank}} r_i^{\text{rank}} + \lambda_{\text{corr}} r_i^{\text{corr}} + \lambda_{\text{fmt}} r_i^{\text{fmt}}, \tag{11}$$

with the outer objective

$$\mathcal{J}(\theta) = \mathbb{E}_{\pi_\theta}\left[\frac{1}{n} \sum_{i=1}^{n} \tilde{r}_i\right]. \tag{12}$$

Here, r_i^{task} is the benchmark-specific correctness reward, r_i^{rank} and r_i^{corr} are rollout-level shaping terms induced by the two RAC objectives (Eqs. 6 and 9), and r_i^{fmt} is a lightweight regularizer that stabilizes the structured output format and can be decayed during training. The outer policy update is carried out with GRPO. RAC therefore leaves the underlying RL optimizer unchanged, and it only changes the reward signal so that confidence behavior becomes part of the learning objective.

Algorithm 1 Ranking-Aware Multimodal RL

Require: Multimodal policy π_θ, clean prompt $x^{(0)}$, corruption operator T_s, severity distribution $p(s)$, rollout count n.

Sample corruption severity $s \sim p(s)$.
Form corrupted input $x^{(s)} = T_s(x^{(0)})$.
for $k = 1$ **to** n **do**
    Sample clean rollout $(y_k^{(0)}, c_k^{(0)}) \sim \pi_\theta(\cdot \mid x^{(0)})$.
    Sample corrupted rollout $(y_k^{(s)}, c_k^{(s)}) \sim \pi_\theta(\cdot \mid x^{(s)})$.
    Obtain verifier scores $a_k^{(0)}, a_k^{(s)} \in \{0, 1\}$.
    Obtain format indicators $f_k^{(0)}, f_k^{(s)} \in \{0, 1\}$.
end for
Find valid clean comparisons $\mathcal{Q}(x^{(0)}) = \{(i, j) : a_i^{(0)} > a_j^{(0)}, f_i^{(0)} = f_j^{(0)} = 1\}$.
Find valid corrupted comparisons $\mathcal{Q}(x^{(s)})$ analogously.
Evaluate ranking group losses $L_{\text{rank}}(x^{(0)})$ and $L_{\text{rank}}(x^{(s)})$ (Eq. 6).
Evaluate pairwise corruption loss $L_{\text{corr}}(x^{(0)}, x^{(s)})$ (Eq. 9).
Compute shaped sequence rewards $\tilde{r}_k^{(0)}$ and $\tilde{r}_k^{(s)}$ (Eq. 11) using task returns and calibration terms.
Update policy π_θ using GRPO over the combined mini-batch, with advantages normalized separately within each prompt-level group.

4 Experiment

This section presents our experimental setup, main results, and ablation studies. We evaluate RAC across four multimodal backbones, six reasoning benchmarks, and five corruption severity levels to assess whether training-time confidence supervision improves calibration without sacrificing accuracy. We further ablate the effect of training corruption severity on calibration generalization and conduct a sensitivity analysis on the loss hyperparameters to validate key design choices.

4.1 Experimental Setup

Backbones. We evaluate four open multimodal checkpoints that span a representative range of model scales and architectures: Qwen2.5-VL-3B-Instruct and Qwen2.5-VL-7B-Instruct (Bai et al., 2025), Qwen3.5-4B, and InternVL-3.5-8B (Zhu et al., 2025). These models are strong enough to expose nontrivial confidence failures while remaining tractable for RL post-training.

Training mixture. The training set is formed by merging appropriate subsets of M3CoT (Chen et al., 2024a), Geometry3K (Lu et al., 2021), and ScienceQA (Lu et al., 2022), retaining only multiple-choice items. Each training pair consists of a corrupted branch A and a clean branch B. Branch A is sampled from a mixed-corruption stream in which 50% of samples are drawn at severity T0.2, 40% at T0.4, and 10% at T0.6, while branch B remains clean. This severity mixture is designed to provide meaningful directional contrast between branches while preserving sufficient visual content for the model to reason about. Image corruptions follow the ImageNet-C protocol (Hendrycks & Dietterich, 2019), covering 8 diverse noise types (detailed in Appendix E), with the specific operator chosen uniformly at random for each sample.

Hyperparameters. All experiments are implemented using the VERL framework (Sheng et al., 2025). We sample $n = 8$ rollouts per prompt with a global batch size of 512 sequences. We use a structured chat template that elicits a verbalized scalar confidence alongside the reasoning chain and final answer (Appendix C). Responses are decoded with a sampling temperature of 0.6, a top-p threshold of 0.95, and a top-k cutoff of 20, with a maximum response length of 3,072 tokens. The policy is optimized with a learning rate of 1×10^{-6} under cosine decay. The ranking margin is set to $m_{\text{rank}} = 0.05$, and reward weights in Eq. 11 are $\lambda_{\text{rank}} = 0.2$ and $\lambda_{\text{corr}} = 0.3$. The format weight λ_{fmt} is initialized at 1.0 and linearly decayed to 0.3 to prevent format supervision from competing with calibration signals at convergence.

Evaluation setup. We evaluate on six multiple-choice benchmarks spanning mathematics, science, and broader multi-discipline domains: M3CoT (Chen et al., 2024a), ScienceQA (Lu et al., 2022),

Table 1: Six-benchmark macro averages: Clean (T0.0), Mild (T0.2–T0.4), Severe (T0.6–T1.0). All models use GRPO. **Bold**/underline: best/second-best Avg per backbone.

Variant	Accuracy ↑				ECE ↓				Brier ↓			
	Clean	Mild	Severe	Avg	Clean	Mild	Severe	Avg	Clean	Mild	Severe	Avg
InternVL-3.5-8B												
Vanilla-RL	0.437	0.441	0.436	0.438	0.354	0.359	0.364	0.359	0.384	0.388	0.388	0.386
+Pair	0.529	0.521	0.509	0.520	0.350	0.356	0.369	0.358	0.350	0.355	0.368	0.358
+Rank	0.566	0.560	0.545	**0.557**	0.320	0.337	0.350	0.336	0.322	0.338	0.349	0.336
+Pair+Rank	0.552	0.546	0.533	0.544	0.303	0.309	0.320	**0.311**	0.307	0.315	0.325	**0.316**
Qwen2.5-VL-3B-Instruct												
Vanilla-RL	0.527	0.516	0.500	0.514	0.319	0.328	0.342	0.329	0.341	0.348	0.359	0.349
+Pair	0.519	0.511	0.495	0.508	0.295	0.304	0.317	0.305	0.316	0.326	0.335	0.326
+Rank	0.557	0.546	0.530	0.544	0.329	0.338	0.352	0.340	0.351	0.358	0.369	0.359
+Pair+Rank	0.590	0.582	0.568	**0.580**	0.287	0.294	0.302	**0.294**	0.308	0.315	0.319	**0.314**
Qwen2.5-VL-7B-Instruct												
Vanilla-RL	0.624	0.623	0.606	0.617	0.316	0.317	0.332	0.322	0.328	0.330	0.344	0.334
+Pair	0.617	0.612	0.598	0.609	0.300	0.311	0.322	0.311	0.314	0.322	0.334	0.324
+Rank	0.640	0.640	0.625	**0.635**	0.325	0.330	0.344	0.333	0.337	0.343	0.356	0.345
+Pair+Rank	0.628	0.616	0.601	0.615	0.286	0.303	0.315	**0.302**	0.299	0.315	0.330	**0.315**
RLCR	0.632	0.597	0.585	0.605	0.283	0.307	0.314	**0.302**	0.298	0.329	0.336	0.322
Qwen3.5-4B												
Vanilla-RL	0.613	0.598	0.581	0.597	0.355	0.362	0.369	0.362	0.360	0.368	0.380	0.369
+Pair	0.616	0.599	0.591	0.602	0.341	0.270	0.251	0.287	0.344	0.301	0.303	0.316
+Rank	0.609	0.595	0.576	0.593	0.377	0.384	0.405	0.389	0.377	0.384	0.405	0.389
+Pair+Rank	0.618	0.601	0.591	**0.603**	0.304	0.228	0.214	**0.248**	0.322	0.277	0.276	**0.292**
RLCR	0.588	0.586	0.561	0.578	0.253	0.251	0.257	0.254	0.290	0.297	0.297	0.294

Table 2: Post-hoc calibration on Qwen3.5-4B. TS: temperature scaling; Platt: Platt scaling. Accuracy is unchanged by calibration. **Bold**/underline: best/second-best values.

Policy	Accuracy ↑	Uncalibrated		TS		Platt	
		ECE ↓	Brier ↓	ECE ↓	Brier ↓	ECE ↓	Brier ↓
Vanilla-RL	0.597	0.362	0.369	0.063	0.242	0.062	0.243
RAC	**0.603**	0.248	0.292	0.079	0.242	**0.017**	**0.238**

MathVerse (Zhang et al., 2025), MMMU (Yue et al., 2024), We-Math (Qiao et al., 2025), and MathVision (Wang et al., 2024a), retaining only multiple-choice items to enable unambiguous answer verification and reliable confidence elicitation via a fixed response schema. Corruption severity $T \in \{0.0, 0.2, 0.4, 0.6, 0.8, 1.0\}$ ranges from clean (T0.0) to maximally degraded (T1.0), with the specific corruption operator drawn uniformly at random from the 8 types (Appendix E), yielding $6 \times 6 = 36$ evaluation subsets in total. We report three metrics macro-averaged across all six benchmarks: Accuracy, Expected Calibration Error (ECE) (Naeini et al., 2015), and Brier score (Brier, 1950). Formal definitions of ECE and Brier score are provided in Appendix F. Results are further aggregated into three severity bands: **Clean** (T0.0), **Mild** (T0.2–T0.4 average), and **Severe** (T0.6–T1.0 average).

4.2 Main Results

Table 1 shows that the ranking-aware loss (+*Rank*) improves average accuracy on most backbones: three of the four evaluated models, with Qwen3.5-4B as the exception. The pairwise corruption loss (+*Pair*) reduces average ECE and Brier score on all four backbones, while its effect on accuracy varies across models. Combining both objectives under +*Pair*+*Rank* improves calibration relative to Vanilla-RL on all four backbones and improves accuracy on three of them. The combined method achieves the lowest average Brier score on all four backbones and the lowest reported ECE, tying with RLCR (Damani et al., 2025) on Qwen2.5-VL-7B at the displayed precision.

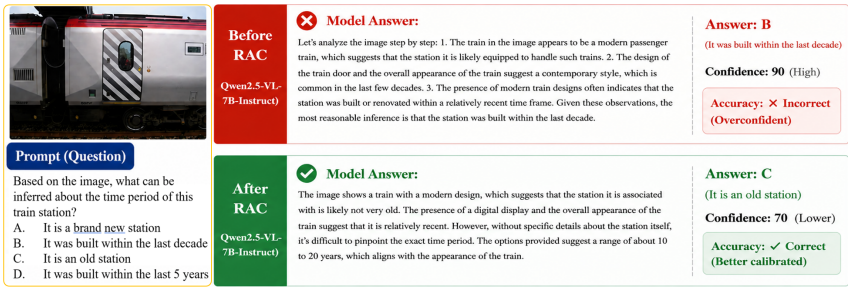
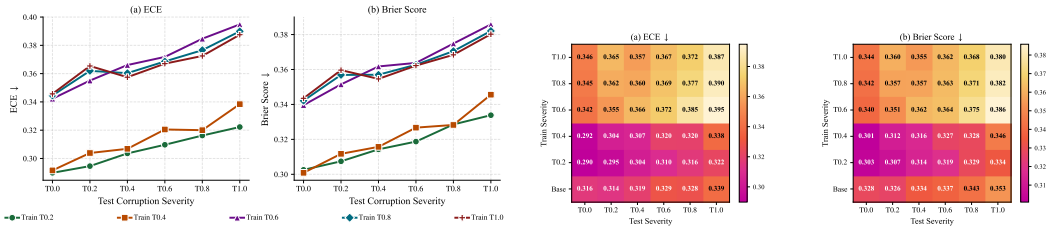


Figure 2: Qualitative case study on Qwen2.5-VL-7B-Instruct. Before RAC, the model produces an incorrect answer (B) with high confidence (90), reflecting overconfident reasoning from ambiguous visual evidence. After RAC training, the same model arrives at the correct answer (C) with appropriately reduced confidence (70), demonstrating improved alignment between stated certainty and evidence quality.



(a) Severity-transfer curves.

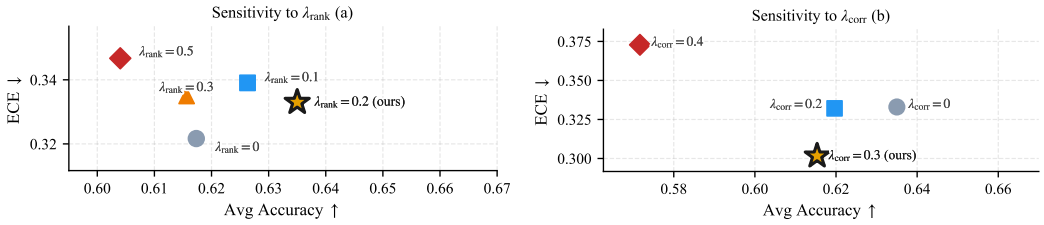
(b) Severity-transfer heatmaps.

Figure 3: **Effect of training corruption severity on calibration transfer** for Qwen2.5-VL-7B-Instruct. Models trained at mild corruption levels (T0.2, T0.4) maintain consistently lower ECE and Brier scores across all test severities, while those trained at T0.6 and above exhibit uniformly degraded calibration. The severity-transfer curves in (a) separate into two distinct clusters at T0.6, a pattern confirmed by the heatmaps in (b).

The accuracy gains from *+Rank* are most pronounced for InternVL-3.5-8B, where average accuracy increases from 0.438 to 0.557, an absolute gain of 11.9 percentage points. Applying the full *+Pair+Rank* framework reduces ECE from 0.359 to 0.311 and Brier from 0.386 to 0.316. On Qwen2.5-VL-7B, *+Rank* achieves the highest accuracy (0.635), while *+Pair+Rank* improves calibration over Vanilla-RL (ECE 0.302 versus 0.322; Brier 0.315 versus 0.334) at a small accuracy cost (0.615 versus 0.617). RAC matches RLCR in reported ECE (both 0.302), with lower Brier score (0.315 versus 0.322) and higher accuracy (0.615 versus 0.605). On Qwen3.5-4B, RAC achieves the best average accuracy, ECE, and Brier score (0.603, 0.248, and 0.292), outperforming RLCR on all three metrics. These results show that the accuracy–calibration trade-off depends on the backbone. Figure 2 provides a qualitative illustration of this behavior. On the same ambiguous visual question, the model assigns high confidence (90%) to an incorrect answer before RAC training, but after training produces the correct answer with appropriately reduced confidence (70%), reflecting the improved alignment between stated certainty and evidence quality that the two calibration losses are designed to instill.

Post-hoc calibration. Table 2 reports five-fold cross-fitting results on Qwen3.5-4B, using the verbalized confidence as input. Temperature scaling and Platt scaling improve calibration for both policies without changing their answers or accuracy. Vanilla-RL with post-hoc calibration already achieves lower ECE and Brier score than uncalibrated RAC. RAC combined with Platt scaling obtains the lowest ECE (0.017) and Brier score (0.238) in this comparison, whereas RAC with temperature scaling has higher ECE than Vanilla-RL with temperature scaling (0.079 versus 0.063). *RAC is complementary to post-hoc calibration, which can further improve its learned confidence estimates.* Appendix B gives the transformations, cross-fitting protocol, and fitted parameters.

Effect of training corruption severity on calibration generalization. To study how training-time corruption strength affects calibration transfer, we train five variants of Qwen2.5-VL-7B-Instruct using different *fixed* corruption levels for branch A: T0.2, T0.4, T0.6, T0.8, and T1.0. Branch B



(a) λ_{rank} sweep (+Rank). (b) λ_{corr} sweep (+Rank+Pair, fixed $\lambda_{\text{rank}} = 0.2$).

Figure 4: Hyperparameter sweeps on Qwen2.5-VL-7B-Instruct (macro-averaged Accuracy and ECE).

remains clean throughout, and all other hyperparameters are held constant. Each variant is then evaluated on all six test severities, reporting macro-averaged Accuracy, ECE, and Brier score across the six benchmarks. Figures 3a and 3b reveal a clear non-monotonic relationship between training corruption strength and calibration quality. Introducing a mild corruption signal consistently benefits the model: the $\mathrm{T}0.2$ variant achieves lower ECE across all test severities and retains the highest task accuracy among all fixed-corruption configurations. Beyond this moderate level, however, further increasing training corruption proves detrimental. The $\mathrm{T}0.6$, $\mathrm{T}0.8$, and $\mathrm{T}1.0$ variants all exhibit uniformly degraded calibration and weaker accuracy, indicating that excessively heavy corruption destabilizes the supervision signal rather than improving robustness. These results suggest that an effective training corruption level strikes a balance: strong enough to expose the model to degraded visual evidence and encourage appropriate confidence attenuation, yet moderate enough to preserve sufficient visual content and avoid inducing systematically under-confident predictions.

Hyperparameter sensitivity. We evaluate the sensitivity of both reward weights by sweeping each over a range of values while holding all other hyperparameters fixed, using Qwen2.5-VL-7B-Instruct throughout. For λ_{rank}, we train five *+Rank* variants with $\lambda_{\text{rank}} \in \{0, 0.1, 0.2, 0.3, 0.5\}$, where $\lambda_{\text{rank}} = 0$ reduces to Vanilla RL. For λ_{corr}, we train four *+Rank+Pair* variants with $\lambda_{\text{corr}} \in \{0, 0.2, 0.3, 0.4\}$, where $\lambda_{\text{corr}} = 0$ reduces to *+Rank* alone. Figure 4 plots macro-averaged accuracy against ECE for each configuration, with the lower-right region indicating better joint performance. For λ_{rank} (Figure 4a), the setting $\lambda_{\text{rank}} = 0.2$ achieves the highest accuracy (0.635) with ECE remaining well-controlled at 0.333. Larger values degrade both metrics, suggesting that an excessively strong ranking penalty impairs rather than reinforces the confidence-ordering objective. For λ_{corr} (Figure 4b), increasing the weight from 0 to 0.3 reduces ECE from 0.333 to 0.302 at a modest accuracy cost, identifying $\lambda_{\text{corr}} = 0.3$ as the optimal setting. At $\lambda_{\text{corr}} = 0.4$, accuracy drops sharply to 0.572 and ECE rises again, indicating that an excessively large corruption penalty destabilizes training. Together, these sweeps confirm that $\lambda_{\text{rank}} = 0.2$ and $\lambda_{\text{corr}} = 0.3$ consistently identify the best accuracy-calibration trade-off across both reward dimensions.

5 Conclusion

We presented RAC, a training-time calibration framework that addresses a structural gap in multimodal RL post-training: existing outcome-reward objectives optimize for answer correctness but provide no signal governing how confidence should track reasoning quality or visual evidence quality. RAC closes this gap with two relative supervision signals arising naturally from group-based RL sampling. The ranking-aware group loss enforces that correct reasoning attracts higher confidence than incorrect reasoning within the same prompt, with observed accuracy gains on most evaluated backbones. The clean–corrupted pairwise loss enforces confidence attenuation as visual evidence weakens, reducing ECE and Brier score. Together, their combination achieves improvements of up to 20.9% in Brier and 31.5% in ECE relative to Vanilla RL, with both objectives requiring no external annotations and no changes to the underlying optimizer.

6 AI USE STATEMENT

In this work, we used generative AI tools for assisting in the writing of proofs, cleaning and reformatting dataset and grammatical refinement toward this manuscript. We have not used generative AI tools for generating synthetic data sets, proposing or refining hypotheses, helping develop theoretical models or conceptual frameworks, formulating mathematical claims, providing critical ingredients for proving mathematical claims, designing or providing feedback on research methodology or experiments, implementing methods, assisting with translation, supporting qualitative and thematic data analysis and interpreting results. Additionally, we used generative AI tools for editing software code (debugging), summarizing or analyzing existing literature, discovering research topics or identifying gaps, brainstorming, sourcing/searching for information, editing a research paper to improve readability, identifying relevant literature, formatting references. We have reviewed all AI-assisted work. To be specific, LLM-generated code was verified and tested for correctness and during grammatically refining this manuscript using AI tools, we manually checked contents generated by AI and substantially changed AI-written text. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

7 ETHICS STATEMENT

The potential negative social impacts of our method are minor, and align with those typically associated with general LLM reasoning technologies. We emphasise the importance of adhering to the principles of fair and safe deployment of LLM systems, including careful consideration of the training data, transparency about model capabilities and limitations, and ongoing monitoring for unintended consequences.

8 REPRODUCIBILITY STATEMENT

The paper fully discloses all information required to reproduce its main experimental results that underpin the core claims and conclusions of the work. For the proposed method and all baselines, the paper (with the Appendix) provides complete and clear descriptions of the experimental design, evaluation protocols, and all necessary implementation details to enable independent replication of the results, in full compliance with ICLR reproducibility requirements.

REFERENCES

Arash Ahmadian, Chris Cremer, Matthias Gallé, Marzieh Fadaee, Julia Kreutzer, Olivier Pietquin, Ahmet Üstün, and Sara Hooker. Back to basics: Revisiting REINFORCE-style optimization for learning from human feedback in LLMs. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 12248–12267. Association for Computational Linguistics, August 2024. doi: 10.18653/v1/2024.acl-long.662. URL https://aclanthology.org/2024.acl-long.662/.

Deepseek AI. Deepseek-r1 incentivizes reasoning in llms through reinforcement learning. *Nature*, 645(8081):633–638, September 2025. ISSN 1476-4687. doi: 10.1038/s41586-025-09422-z. URL http://dx.doi.org/10.1038/s41586-025-09422-z.

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob L Menick, Sebastian Borgeaud, Andy Brock, Aida Nematzadeh, Sahand Sharifzadeh, Mikoł aj Bińkowski, Ricardo Barreira, Oriol Vinyals, Andrew Zisserman, and Karén Simonyan. Flamingo: a visual language model for few-shot learning. In *Advances in Neural Information Processing Systems*, volume 35, pp. 23716–23736, 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/960a172bc7fbf0177ccccbb411a7d800-Paper-Conference.pdf.

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond, 2023. URL https://arxiv.org/abs/2308.12966.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-vl technical report, 2025. URL https://arxiv.org/abs/2502.13923.

Glenn W. Brier. Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, 78(1):1–3, 1950.

Ji Young Byun, Young-Jin Park, Jean-Philippe Corbeil, and Asma Ben Abacha. Overconfidence and calibration in medical vqa: Empirical findings and hallucination-aware mitigation, 2026. URL https://arxiv.org/abs/2604.02543.

Qiguang Chen, Libo Qin, Jin Zhang, Zhi Chen, Xiao Xu, and Wanxiang Che. M^3CoT: 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)*, pp. 8199–8221, 2024a. URL https://aclanthology.org/2024.acl-long.446/.

Xuweiyi Chen, Ziqiao Ma, Xuejun Zhang, Sihan Xu, Jianing Yang, David F. Fouhey, Joyce Chai, and Shengyi Qian. Multi-object hallucination in vision language models. In *Advances in Neural Information Processing Systems*, volume 37, pp. 44393–44418, 2024b. doi: 10.52202/079017-1409. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/4ea4a1ea4d9ff273688c8e92bd087112-Paper-Conference.pdf.

Mehul Damani, Isha Puri, Stewart Slocum, Idan Shenfeld, Leshem Choshen, Yoon Kim, and Jacob Andreas. Beyond binary rewards: Training lms to reason about their uncertainty, 2025. URL https://arxiv.org/abs/2507.16806.

Yunkai Dang, Yifan Jiang, Yizhu Jiang, Anqi Chen, Wenbin Li, and Yang Gao. Instinct vs. reflection: Unifying token and verbalized confidence in multimodal large models, 2026. URL https://arxiv.org/abs/2604.17274.

Shrey Desai and Greg Durrett. Calibration of pre-trained transformers. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pp. 295–302. Association for Computational Linguistics, November 2020. doi: 10.18653/v1/2020.emnlp-main.21. URL https://aclanthology.org/2020.emnlp-main.21/.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 1321–1330, 06–11 Aug 2017. URL https://proceedings.mlr.press/v70/guo17a.html.

Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=HJz6tiCqYm.

Wenxuan Huang, Bohan Jia, Zijie Zhai, Shaosheng Cao, Zheyu Ye, Fei Zhao, Zhe Xu, Xu Tang, Yao Hu, and Shaohui Lin. Vision-r1: Incentivizing reasoning capability in multimodal large language models, 2025. URL https://arxiv.org/abs/2503.06749.

Ziwei Ji, Lei Yu, Yeskendir Koishekenov, Yejin Bang, Anthony Hartshorn, Alan Schelten, Cheng Zhang, Pascale Fung, and Nicola Cancedda. Calibrating verbal uncertainty as a linear feature to reduce hallucinations. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 3769–3793, 2025. URL https://aclanthology.org/2025.emnlp-main.187/.

Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, Scott Johnston, Sheer El-Showk, Andy Jones, Nelson Elhage, Tristan Hume, Anna Chen, Yuntao Bai, Sam Bowman, Stanislav Fort, Deep Ganguli, Danny Hernandez, Josh Jacobson, Jackson Kernion, Shauna Kravec, Liane Lovitt, Kamal Ndousse, Catherine Olsson, Sam Ringer, Dario Amodei, Tom Brown, Jack Clark, Nicholas Joseph, Ben Mann, Sam McCandlish, Chris Olah, and Jared Kaplan. Language models (mostly) know what they know, 2022. URL https://arxiv.org/abs/2207.05221.

Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. In *Advances in Neural Information Processing Systems*, volume 33, 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/hash/d89a66c7c80a29b1bdbab0f2a1a94af8-Abstract.html.

Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation. In *International Conference on Learning Representations (ICLR)*, 2023. URL https://openreview.net/forum?id=VD-AYtP0dve.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: bootstrapping language-image pre-training with frozen image encoders and large language models. In *Proceedings of the 40th International Conference on Machine Learning*, 2023a.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 292–305, December 2023b. doi: 10.18653/v1/2023.emnlp-main.20. URL https://aclanthology.org/2023.emnlp-main.20/.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *Advances in Neural Information Processing Systems*, volume 36, pp. 34892–34916, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/6dcf277ea32ce3288914faf369fe6de0-Paper-Conference.pdf.

Ziyu Liu, Zeyi Sun, Yuhang Zang, Xiaoyi Dong, Yuhang Cao, Haodong Duan, Dahua Lin, and Jiaqi Wang. Visual-rft: Visual reinforcement fine-tuning, 2025. URL https://arxiv.org/abs/2503.01785.

Pan Lu, Ran Gong, Shibiao Jiang, Liang Qiu, Siyuan Huang, Xiaodan Liang, and Song-Chun Zhu. Inter-GPS: Interpretable geometry problem solving with formal language and symbolic reasoning. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 6774–6786, 2021. doi: 10.18653/v1/2021.acl-long.528. URL https://aclanthology.org/2021.acl-long.528/.

Pan Lu, Swaroop Mishra, Tanglin Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. Learn to explain: Multimodal reasoning via thought chains for science question answering. In *Advances in Neural Information Processing Systems*, volume 35, pp. 2507–2521, 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/11332b6b6cf4485b84afadb1352d3a9a-Paper-Conference.pdf.

Matthias Minderer, Josip Djolonga, Rob Romijnders, Frances Hubis, Xiaohua Zhai, Neil Houlsby, Dustin Tran, and Mario Lucic. Revisiting the calibration of modern neural networks. In *Advances in Neural Information Processing Systems*, volume 34, pp. 15682–15694, 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/file/8420d359404024567b5aefda1231af24-Paper.pdf.

Mahdi Pakdaman Naeini, Gregory Cooper, and Milos Hauskrecht. Obtaining well calibrated probabilities using Bayesian binning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 29, 2015. doi: 10.1609/aaai.v29i1.9602. URL https://ojs.aaai.org/index.php/AAAI/article/view/9602.

Alexandru Niculescu-Mizil and Rich Caruana. Predicting good probabilities with supervised learning. In *Proceedings of the 22nd International Conference on Machine Learning*, pp. 625–632, 2005. ISBN 1595931805. doi: 10.1145/1102351.1102430. URL https://doi.org/10.1145/1102351.1102430.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*, volume 35, pp. 27730–27744, 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/b1efde53be364a73914f58805a001731-Paper-Conference.pdf.

Yaniv Ovadia, Emily Fertig, Jie Ren, Zachary Nado, D. Sculley, Sebastian Nowozin, Joshua Dillon, Balaji Lakshminarayanan, and Jasper Snoek. Can you trust your model's uncertainty? evaluating predictive uncertainty under dataset shift. In *Advances in Neural Information Processing Systems*, volume 32, 2019. URL https://proceedings.neurips.cc/paper_files/paper/2019/file/8558cb408c1d76621371888657d2eb1d-Paper.pdf.

Runqi Qiao, Qiuna Tan, Guanting Dong, Minhui Wu, Chong Sun, Xiaoshuai Song, Jiapeng Wang, Zhuoma GongQue, Shanglin Lei, YiFan Zhang, Zhe Wei, Miaoxuan Zhang, Runfeng Qiao, Xiao Zong, Yida Xu, Peiqing Yang, Zhimin Bao, Muxi Diao, Chen Li, and Honggang Zhang. We-math: Does your large multimodal model achieve human-like mathematical reasoning? In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 20023–20070, 2025. doi: 10.18653/v1/2025.acl-long.983. URL https://aclanthology.org/2025.acl-long.983/.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139, pp. 8748–8763, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/radford21a.html.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. In *Advances in Neural Information Processing Systems*, volume 36, pp. 53728–53741, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/a85b405ed65c6477a4fe8302b5e06ce7-Paper-Conference.pdf.

Rohit Saxena, Alessandro Suglia, and Pasquale Minervini. Vlm-robustbench: A comprehensive benchmark for robustness of vision-language models, 2026. URL https://arxiv.org/abs/2603.06148.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms, 2017. URL https://arxiv.org/abs/1707.06347.

Guangming Sheng, Chi Zhang, Zilingfeng Ye, Xibin Wu, Wang Zhang, Ru Zhang, Yanghua Peng, Haibin Lin, and Chuan Wu. Hybridflow: A flexible and efficient rlhf framework. In *Proceedings of the Twentieth European Conference on Computer Systems*, pp. 1279–1297. ACM, March 2025. doi: 10.1145/3689031.3696075. URL http://dx.doi.org/10.1145/3689031.3696075.

Elias Stengel-Eskin, Peter Hase, and Mohit Bansal. LACIE: Listener-aware finetuning for calibration in large language models. In *Advances in Neural Information Processing Systems*, volume 37, pp. 43080–43106, 2024. URL https://papers.nips.cc/paper_files/paper/2024/file/4b8eaf3bcdc105423a972ed90eb07217-Paper-Conference.pdf.

Xiangjie Sui, Songyang Li, Hanwei Zhu, Baoliang Chen, Yuming Fang, and Xin Sun. Benchmarking corruption robustness of lvlms: A discriminative benchmark and robustness alignment metric, 2025. URL https://arxiv.org/abs/2511.19032v1.

DeepSeek-AI Team. Deepseekmath: Pushing the limits of mathematical reasoning in open language models, 2024. URL https://arxiv.org/abs/2402.03300.

Katherine Tian, Eric Mitchell, Allan Zhou, Archit Sharma, Rafael Rafailov, Huaxiu Yao, Chelsea Finn, and Christopher Manning. Just ask for calibration: Strategies for eliciting calibrated confidence scores from language models fine-tuned with human feedback. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 5433–5442, 2023. URL https://aclanthology.org/2023.emnlp-main.330/.

Jiahao Wang, Weiye Xu, Aijun Yang, Wengang Zhou, Lewei Lu, Houqiang Li, Xiaohua Wang, and Jinguo Zhu. Enhancing the outcome reward-based rl training of mllms with self-consistency sampling. In *Advances in Neural Information Processing Systems*, volume 38, pp. 132018–132052, 2025a. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/befcd01725459b183bfa92058315c379-Paper-Conference.pdf.

Ke Wang, Junting Pan, Weikang Shi, Zimu Lu, Houxing Ren, Aojun Zhou, Mingjie Zhan, and Hongsheng Li. Measuring multimodal mathematical reasoning with math-vision dataset. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 95095–95169, 2024a. doi: 10.52202/079017-3014. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/ad0edc7d5fa1a783f063646968b7315b-Paper-Datasets_and_Benchmarks_Track.pdf.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution, 2024b. URL https://arxiv.org/abs/2409.12191.

Weiyun Wang, Zhe Chen, Wenhai Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Jinguo Zhu, Xizhou Zhu, Lewei Lu, Yu Qiao, and Jifeng Dai. Enhancing the reasoning ability of multimodal large language models via mixed preference optimization, 2025b. URL https://arxiv.org/abs/2411.10442.

Wenyi Xiao, Xinchi Xu, and Leilei Gan. Vl-calibration: Decoupled confidence calibration for large vision-language models reasoning, 2026. URL https://arxiv.org/abs/2604.09529.

Johnathan Xie, Annie S. Chen, Yoonho Lee, Eric Mitchell, and Chelsea Finn. Calibrating language models with adaptive temperature scaling. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 18128–18138, 2024. URL https://aclanthology.org/2024.emnlp-main.1007/.

Miao Xiong, Zhiyuan Hu, Xinyang Lu, Yifei Li, Jie Fu, Junxian He, and Bryan Hooi. Can llms express their uncertainty? an empirical evaluation of confidence elicitation in llms. In *International Conference on Learning Representations*, volume 2024, pp. 23650–23678, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/file/6733cf15e10e2cd1d59af033c3bb8507-Paper-Conference.pdf.

Dongkeun Yoon, Seungone Kim, Sohee Yang, Sunkyoung Kim, Soyeon Kim, Yongil Kim, Eunbi Choi, Yireun Kim, and Minjoon Seo. Reasoning models better express their confidence, 2025. URL https://arxiv.org/abs/2505.14489.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, Cong Wei, Botao Yu, Ruibin Yuan, Renliang Sun, Ming Yin, Boyuan Zheng, Zhenzhu Yang, Yibo Liu, Wenhao Huang, Huan Sun, Yu Su, and Wenhu Chen. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9556–9567, 2024. doi: 10.1109/CVPR52733.2024.00913. URL https://openaccess.thecvf.com/content/CVPR2024/html/Yue_MMMU_A_Massive_Multi-discipline_Multimodal_Understanding_and_Reasoning_Benchmark_for_CVPR_2024_paper.html.

Mozhi Zhang, Mianqiu Huang, Rundong Shi, Linsen Guo, Chong Peng, Peng Yan, Yaqian Zhou, and Xipeng Qiu. Calibrating the confidence of large language models by eliciting fidelity. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 2959–2979, 2024. URL https://aclanthology.org/2024.emnlp-main.173/.

Renrui Zhang, Dongzhi Jiang, Yichi Zhang, Haokun Lin, Ziyu Guo, Pengshuo Qiu, Aojun Zhou, Pan Lu, Kai-Wei Chang, Yu Qiao, Peng Gao, and Hongsheng Li. Mathverse: Does your multi-modal llm truly see the diagrams in visual math problems? In *Computer Vision – ECCV 2024*, volume 15066 of *Lecture Notes in Computer Science*, pp. 169–186, 2025. doi: 10.1007/978-3-031-73242-3_10. URL https://link.springer.com/chapter/10.1007/978-3-031-73242-3_10.

Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, Zhangwei Gao, Erfei Cui, Xuehui Wang, Yue Cao, Yangzhou Liu, Xingguang Wei, Hongjie Zhang, Haomin Wang, Weiye Xu, Hao Li, Jiahao Wang, Nianchen Deng, Songze Li, Yinan He, Tan Jiang, Jiapeng Luo, Yi Wang, Conghui He, Botian Shi, Xingcheng Zhang, Wenqi Shao, Junjun He, Yingtong Xiong, Wenwen Qu, Peng Sun, Penglong Jiao, Han Lv, Lijun Wu, Kaipeng Zhang, Huipeng Deng, Jiaye Ge, Kai Chen, Limin Wang, Min

Dou, Lewei Lu, Xizhou Zhu, Tong Lu, Dahua Lin, Yu Qiao, Jifeng Dai, and Wenhai Wang. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models, 2025. URL https://arxiv.org/abs/2504.10479.

APPENDIX CONTENTS

A Additional Related Works **17**

B Post-hoc Calibration **17**

C Chat Template **18**

D Details of Models and Datasets **19**

E Image Corruptions and Examples **19**

F Evaluation Metrics **20**

G Compute Resources **24**

H Discussion **24**

A Additional Related Works

Reinforcement learning for alignment. Reinforcement learning has increasingly become the standard for post-training large language and multimodal models. Early human-feedback alignment relied on Proximal Policy Optimization (PPO) (Schulman et al., 2017; Ouyang et al., 2022) to match human preferences, while DPO (Rafailov et al., 2023) streamlined this by directly optimizing from offline preference pairs. As the field shifted toward eliciting complex reasoning, the alignment objective evolved from mimicking human style to optimizing verifiable correctness. Outcome-based RL has thus gained prominence for reasoning tasks (Team, 2024; AI, 2025; Wang et al., 2025b), training models against rule-based verifiers to develop long chain-of-thought logic and autonomous self-correction. In multimodal models, this reasoning search is even more demanding, as the policy must additionally align linguistic logic with visual perception. To reduce the memory overhead of critic networks, GRPO estimates baselines directly from multiple rollouts of the same prompt (Team, 2024) and has since been adapted for visual reasoning tasks (Wang et al., 2025a; Liu et al., 2025; Huang et al., 2025).

Confidence calibration and visual shift. Calibration research examines whether a model's stated confidence reflects its true probability of being correct (Niculescu-Mizil & Caruana, 2005; Guo et al., 2017; Brier, 1950). A well-established finding is that calibration degrades sharply under distribution shift and common corruptions (Ovadia et al., 2019; Minderer et al., 2021). Within generative models, the evaluation paradigm has evolved from softmax logits to verbalized confidence and semantic uncertainty estimates (Desai & Durrett, 2020; Kadavath et al., 2022; Kuhn et al., 2023). For multimodal models, the failure mode is more severe: visual anomalies cascade into hallucinated reasoning, whereby models assign high expressed certainty to incorrect answers generated from degraded inputs (Li et al., 2023b; Chen et al., 2024b; Dang et al., 2026; Byun et al., 2026). Post-hoc methods such as temperature scaling and Bayesian binning offer partial relief (Guo et al., 2017; Naeini et al., 2015), and recent work specifically targets VLM calibration via decoupled rescaling (Xiao et al., 2026). However, all such approaches operate on a frozen model and cannot alter what the policy has learned to be certain about.

Calibration and uncertainty in LLMs. A growing body of work examines whether large language models can reliably express their own uncertainty. On the post-hoc side, temperature-based recalibration of RLHF-tuned model probabilities (Kadavath et al., 2022; Tian et al., 2023; Xie et al., 2024) and verbalized confidence elicitation (Xiong et al., 2024; Zhang et al., 2024) have been shown to partially recover calibration that post-training degrades, though all such approaches operate on a frozen model. Training-time interventions go further by making calibration a direct learning signal, through representation editing (Ji et al., 2025), listener-aware fine-tuning (Stengel-Eskin et al., 2024), and calibration rewards integrated into RL post-training (Damani et al., 2025). Analyses of emergent behavior further suggest that reasoning-oriented RL can improve confidence expression as a byproduct, as models engaging in self-verification and backtracking tend to assign confidence more accurately (Yoon et al., 2025). These studies collectively demonstrate that uncertainty can be elicited, calibrated, and trained in text-only LLMs, but they do not address the multimodal setting in which confidence must additionally respond to the quality of visual evidence.

B Post-hoc Calibration

Confidence transformations. Let $p_i \in [0, 1]$ denote the answer-level verbalized confidence parsed from the `<confidence>` tag. Before taking log odds, we clip the confidence to avoid infinite values at the endpoints:

$$\widetilde{p}_i = \mathrm{clip}(p_i, 10^{-4}, 1 - 10^{-4}), \qquad z_i = \mathrm{logit}(\widetilde{p}_i) = \log \frac{\widetilde{p}_i}{1 - \widetilde{p}_i}. \tag{13}$$

Temperature scaling (Guo et al., 2017) and Platt scaling (Niculescu-Mizil & Caruana, 2005) transform these scalar log odds as

$$p_{i,\mathrm{TS}} = \sigma\left(\frac{z_i}{T_{\mathrm{cal}}}\right), \qquad p_{i,\mathrm{Platt}} = \sigma(a z_i + b), \tag{14}$$

where $\sigma(u) = (1 + \exp(-u))^{-1}$, $T_{\text{cal}} > 0$ is the temperature, and a, b are the Platt slope and intercept. The subscript distinguishes the calibration temperature from corruption severity T. Both methods transform the reported confidence and leave generated answers unchanged, so they do not change accuracy.

Fitting objective and optimization. Both methods use binary correctness labels $y_i \in \{0, 1\}$ and minimize Bernoulli negative log-likelihood (NLL) on the fitting subset $\mathcal{S}$:

$$\mathcal{L}_{\mathcal{S}}(\widehat{p}) = -\frac{1}{|\mathcal{S}|} \sum_{i \in \mathcal{S}} \left[y_i \log \widehat{p}_i + (1 - y_i) \log(1 - \widehat{p}_i) \right]. \tag{15}$$

For temperature scaling, SciPy bounded scalar optimization fits $\tau = \log T_{\text{cal}}$ over $\tau \in [-6, 6]$, with $T_{\text{cal}} = e^{\tau}$. For Platt scaling, L-BFGS-B fits (a, b) with bounds $a \in [10^{-6}, 20]$ and $b \in [-20, 20]$, initialized at $(1, 0)$. Raw ECE and Brier scores use the original, unclipped p_i; clipping is used only for the logit transformation and NLL fitting.

Five-fold cross-fitting. We fit calibrators separately for Vanilla-RL and RAC on Qwen3.5-4B. For each held-out fold $k \in \{0, 1, 2, 3, 4\}$, we fit $T_{\text{cal}}^{(k)}$ and $(a^{(k)}, b^{(k)})$ using confidence–correctness pairs from the other four folds, then apply each fitted transformation only to fold k. Repeating this process yields out-of-fold calibrated confidences for every example. Thus, each method has its own parameters for each policy and fold, rather than a single parameter set shared across all five folds. Table 3 lists the fitted parameters used for Table 2.

Table 3: Fold-specific post-hoc calibration parameters for Qwen3.5-4B. Each row gives parameters fitted on the other four folds and applied to the indicated held-out fold.

Policy	Fold	T_{cal}	Platt a	Platt b
Vanilla-RL	0	20.3938	0.130392	-0.750755
	1	31.4415	0.000001	0.295278
	2	22.8118	0.126227	-0.760293
	3	18.4202	0.134630	-0.741362
	4	22.0244	1.579335	-14.129778
RAC (+Pair+Rank)	0	11.6270	0.031620	0.399913
	1	10.4729	0.059418	0.262011
	2	11.2117	0.045787	0.325080
	3	9.4188	0.053426	0.383272
	4	8.8848	0.071034	0.297395

C CHAT TEMPLATE

```
"\n Please reason step by step and follow this exact response schema:\n
    "
"<think>your reasoning</think>\n"
"<answer>your final answer</answer>\n"
"<confidence>your confidence</confidence>\n"
"If options are provided, put only the single option letter inside <
    answer>.\n"
"Otherwise, put only the final value or short expression inside <answer
    >.\n"
"Inside <confidence>, output exactly one decimal number between 0.0 and
     1.0.\n"
"Start your response with <think> and do not output any text before <
    think> or after </confidence>."
```

D Details of Models and Datasets

Models. We fine-tune four open-source vision-language models. **Qwen2.5-VL-3B-Instruct** is a 3-billion-parameter instruction-tuned model from the Qwen2.5-VL family (Bai et al., 2025), released under the Apache 2.0 License (`https://huggingface.co/Qwen/Qwen2.5-VL-3B-Instruct`). **Qwen2.5-VL-7B-Instruct** is the 7-billion-parameter counterpart (Bai et al., 2025), also released under the Apache 2.0 License (`https://huggingface.co/Qwen/Qwen2.5-VL-7B-Instruct`). **InternVL2.5-8B** is an 8-billion-parameter model from the InternVL3 family (Zhu et al., 2025), released under the MIT License (`https://huggingface.co/OpenGVLab/InternVL2_5-8B`). **Qwen3.5-4B** is the additional backbone evaluated in Table 1. These four checkpoints were selected to cover a representative range of scales and architectures while remaining tractable for RL post-training.

Training datasets. **M^{3}CoT** (Chen et al., 2024a) is a multi-domain, multi-step, multi-modal chain-of-thought benchmark covering science, mathematics, and commonsense reasoning (MIT License; `https://huggingface.co/datasets/LightChen233/M3CoT`). **Geometry3K** (Lu et al., 2021) is a formal geometry problem set with diagrams and structured annotations (MIT License; `https://huggingface.co/datasets/hiyouga/geometry3k`). **ScienceQA** (Lu et al., 2022) is a multimodal multiple-choice benchmark spanning natural science, language science, and social science topics (MIT License; `https://huggingface.co/datasets/derek-thomas/ScienceQA`). We retain only the multiple-choice items from each source in the training mixture.

Evaluation datasets. **M^{3}CoT** (Chen et al., 2024a) is also used for evaluation on its held-out test split (MIT License; see above for the dataset link). **ScienceQA** (Lu et al., 2022) is also used for evaluation on its held-out test split (MIT License; see above for the dataset link). **MathVerse** (Zhang et al., 2025) tests whether models can truly perceive diagrams in visual math problems (MIT License; `https://huggingface.co/datasets/AI4Math/MathVerse`). **MMMU** (Yue et al., 2024) is a massive multi-discipline multimodal benchmark spanning 30 university subjects and 183 subfields (Apache 2.0; `https://huggingface.co/datasets/MMMU/MMMU`). **We-Math** (Qiao et al., 2025) provides hierarchical visual math problems designed to probe knowledge acquisition and generalization (CC BY-NC 4.0; `https://huggingface.co/datasets/We-Math/We-Math`). **MathVision** (Wang et al., 2024a) is a competition-level mathematical problem set with visual contexts spanning 16 disciplines (CC BY-NC-SA 4.0; `https://huggingface.co/datasets/mathvision-bench/MathVision`).

E Image Corruptions and Examples

Corruption algorithms. We implement 8 visual degradation operators following the ImageNet-C robustness protocol (Hendrycks & Dietterich, 2019), each mapped across 5 discrete severity levels (T0.2 to T1.0). These corruptions are summarized in Table 4.

Importantly, during both training pair construction and evaluation, the specific corruption operator applied to any given image is sampled uniformly at random from these 8 types. Two images at the same severity level (e.g., both at T0.4) may thus be degraded by entirely different operators (e.g., Fog vs. Elastic Transform), preventing the policy from overfitting to any single noise pattern and encouraging robustness to diverse visual degradation.

Dataset degradation ladder. Figures 6, 7, and 8 demonstrate the effects of increasing corruption severity on samples selected from each dataset. We pair clean base images with their respective T0.2 to T1.0 augmented views to form the 36 evaluation subsets.

Table 4: Summary of visual degradation operators and their corresponding severity parameter scales (T0.2 to T1.0).

Corruption Type	**Description**	**Severity Scale** (T0.2 – T1.0)
Gaussian Noise	Additive zero-mean Gaussian noise.	Std. dev. 0.08 →0.38 (rel. to 255).
Shot Noise	Poisson noise mimicking low-light sensor defects.	Scaled photon counts 60 →3.
Impulse Noise	Salt-and-pepper noise affecting random pixels.	Random fraction 3% →27%.
Fog	Additive synthetic diamond-square plasma fractal.	Increasing fog thickness/intensity.
Brightness	HSV space Value-channel manual adjustments.	Uniform offset 0.1 →0.5.
Contrast	Blending toward the image's mean pixel intensity.	Decreasing standard deviation.
Elastic Transform	Localized warping via affine transformations.	Increasing displacement field vectors.
JPEG Compression	Heavy reduction in image storage quality.	`libjpeg` quality 25 →7.

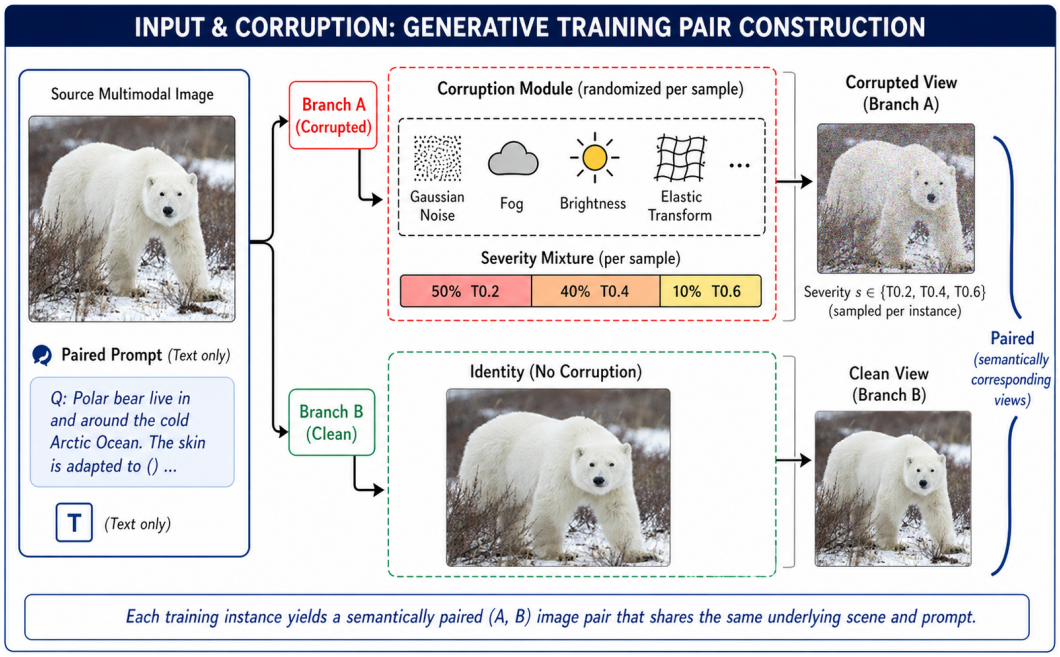


Figure 5: **Generative Training Pair Construction.** The matched multimodal branches underpinning our method: *Branch A (Corrupted)* applies randomly sampled visual perturbation (following the ImageNet-C protocol), juxtaposed against the untouched source in *Branch B (Clean)*. This paired structure provides the clean–corrupted comparisons used in the RAC pairwise loss (Figure 1).

F Evaluation Metrics

We report three primary summary metrics here. Accuracy measures task success and ignores confidence:

$$\text{Acc} = \frac{1}{N}\sum_{i=1}^{N} a_i. \tag{16}$$

The Brier score measures the overall quality of probabilistic confidence by penalizing both overconfident errors and underconfident correct predictions:

$$\text{Brier} = \frac{1}{N}\sum_{i=1}^{N}(c_i - a_i)^2. \tag{17}$$

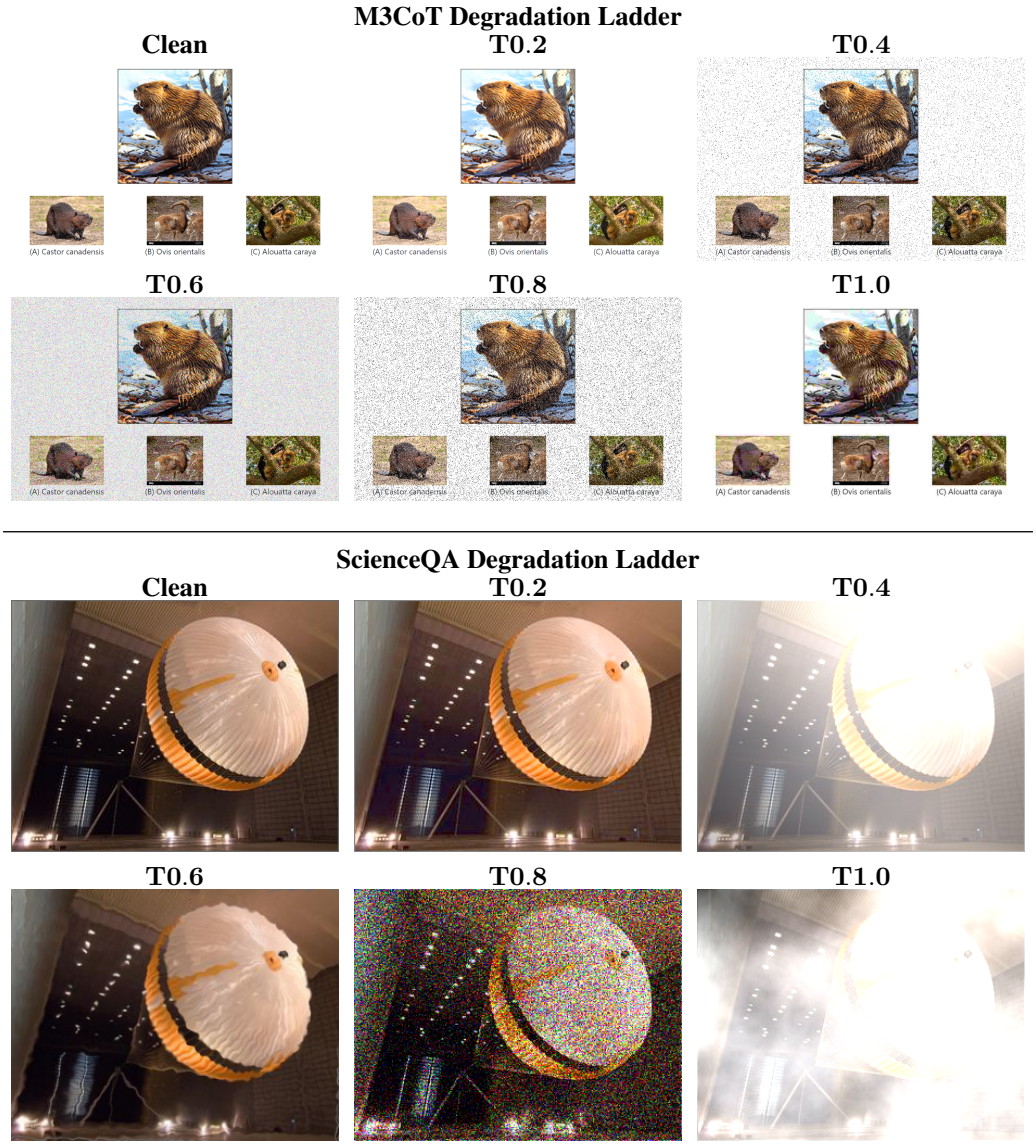


Figure 6: Visual degradation severity ladder from Clean to T1.0 for **M3CoT** (top) and **ScienceQA** (bottom).

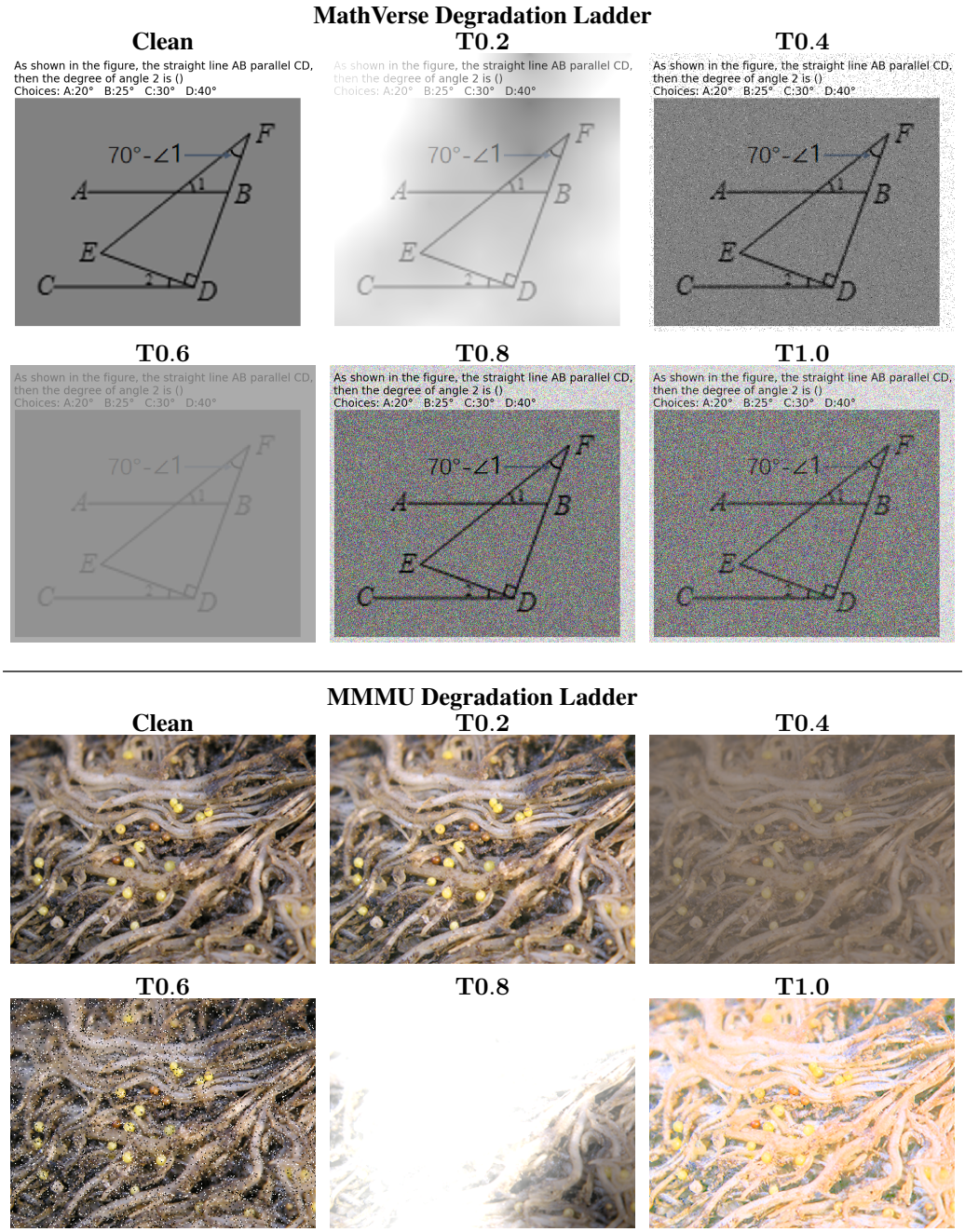


Figure 7: Visual degradation severity ladder from Clean to T1.0 for **MathVerse** (top) and **MMMU** (bottom).

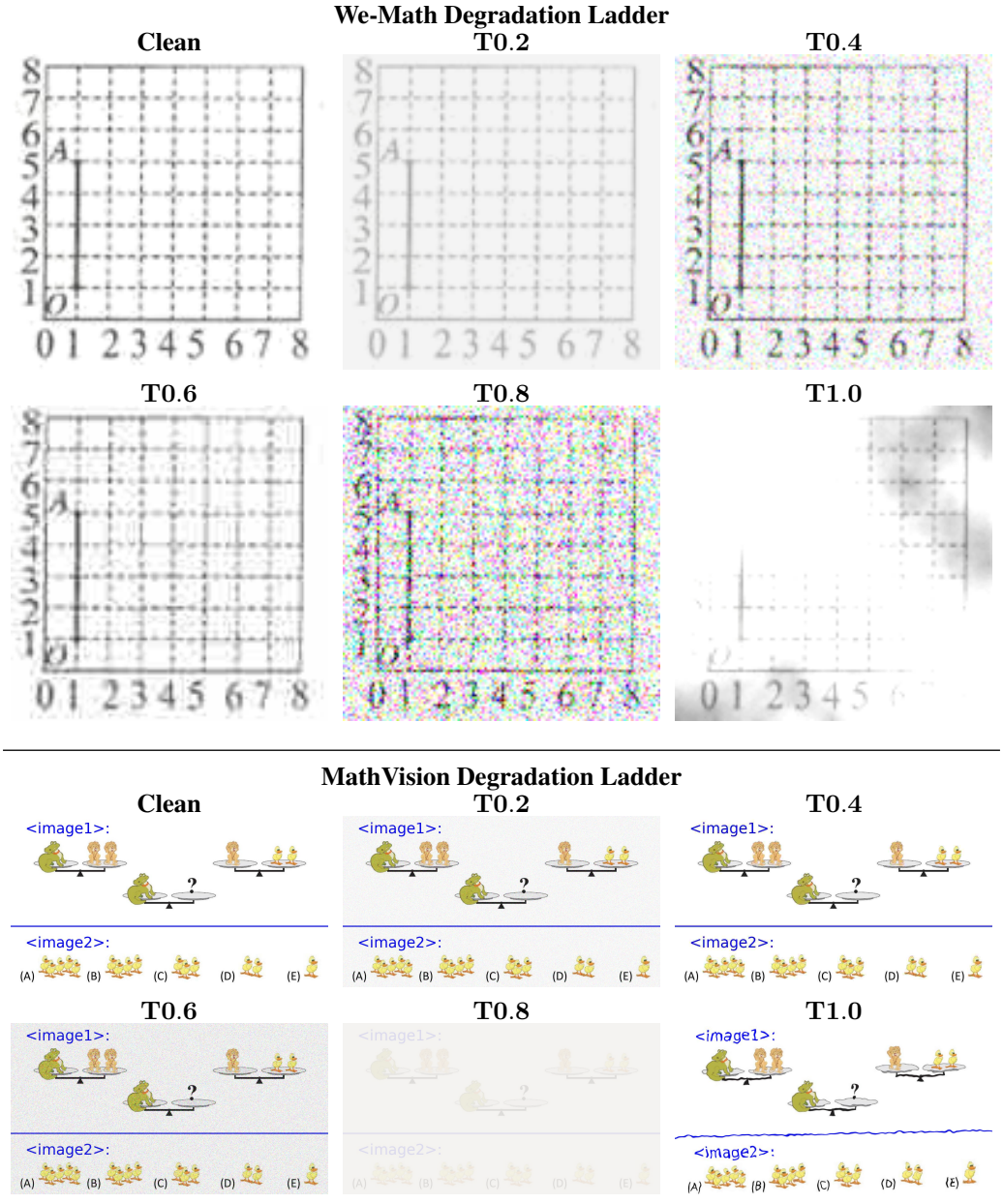


Figure 8: Visual degradation severity ladder from Clean to T1.0 for **We-Math** (top) and **MathVision** (bottom).

Expected calibration error (ECE) measures the average gap between confidence and empirical correctness after binning predictions by confidence:

$$\mathrm{ECE} = \sum_{b=1}^{B} \frac{|\mathcal{I}_b|}{N} \left| \frac{1}{|\mathcal{I}_b|} \sum_{i \in \mathcal{I}_b} a_i - \frac{1}{|\mathcal{I}_b|} \sum_{i \in \mathcal{I}_b} c_i \right|. \tag{18}$$

Here B is the number of confidence bins and $\mathcal{I}_b = \{i : c_i \in ((b-1)/B, b/B]\}$ is the index set of predictions whose reported confidence falls into bin b. The two averages inside the absolute value are therefore the empirical accuracy and the mean reported confidence within that bin. ECE is small only when each confidence range carries the right empirical success rate. In our setting, Acc is higher-is-better, whereas Brier and ECE are lower-is-better, so the three metrics separate answer quality from confidence quality under both clean and corrupted inputs.

G Compute Resources

All experiments were conducted on a cluster of 8 NVIDIA Blackwell-series GPUs. Each training run (one backbone $\times$ one configuration) took approximately 9 hours. Over the course of this project we conducted approximately 40 experimental runs in total, including exploratory runs and ablation studies, amounting to roughly $40 \times 9 \times 8 = 2{,}880$ single-GPU hours in aggregate.

H Discussion

Several directions remain open for future work. Extending RAC to adaptive corruption curricula that progressively diversify severity and operator type could further strengthen calibration transfer to unseen degradations. Furthermore, one limitation is that RAC is currently evaluated on multiple-choice benchmarks, where binary correctness labels provide unambiguous reward signals for confidence calibration. So a longer-term direction is applying RAC to open-ended generation settings, where binary correctness labels are unavailable and confidence must instead be grounded in preference or semantic similarity signals.