CAVE-RL: CONFIDENCE-ADAPTIVE VISUAL EVIDENCE FOR MULTIMODAL REINFORCEMENT LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Reinforcement learning (RL) for multimodal models increasingly relies on learned visual verifiers, but those verifiers are imperfect precisely on the visually ambiguous inputs where policy optimization is most likely to exploit them. We propose CAVE-RL (Confidence-Adaptive Visual Evidence RL), a reward construction that combines sparse exact task feedback with visual proxy rewards whose influence is attenuated by predictive entropy and ensemble disagreement, then anchors policy updates to a frozen reference policy. The method is architecture-agnostic and can be inserted into PPO/GRPO-style post-training whenever an action can be mapped to a visual claim scored by one or more verifiers. We construct 30,555 multimodal image-instruction instances from 1,797 real handwritten-digit images, covering five language-conditioned transformation tasks, held-out prompt paraphrases, visual distribution shifts, 60% corruption in verifier training labels, and five independent random seeds. Only 15% of RL rollouts reveal the exact task outcome. Under this budget, CAVE-RL improves the five-condition robust average by 2.17 pp over SFT and 0.70 pp over sparse-reward PPO, while reducing expected calibration error from 12.49% to 8.34%. It recovers 74% of the robustness gain of an oracle PPO baseline that observes the exact outcome on every rollout.

1 INTRODUCTION

Reinforcement learning with verifiable or preference-based rewards has become a central post-training tool for large language models (Schulman et al., 2017; Ouyang et al., 2022; Shao et al., 2024; DeepSeek-AI, 2025). Multimodal large language models (MLLMs) increasingly inherit the same recipe, but the reward channel is qualitatively different: success depends on both reasoning and visual evidence. Recent work therefore augments outcome rewards with visual perception rewards (Xiao et al., 2026), perception-aware policy objectives (Wang et al., 2026b), task-grounded rewards (Zhang et al., 2026a), active perception (Zheng et al., 2026), dense visual curricula (Feng et al., 2026), and multimodal verifiers (Marsili and Gkioxari, 2026). These advances share a premise that visual feedback can guide reasoning. They also expose a risk: a policy may become better at satisfying an imperfect visual proxy without becoming better at the underlying task. Recent measurements of multimodal reward hacking make this failure mode explicit (Yao et al., 2026).

We study a simple question: *how should a multimodal policy use a visual proxy reward when the proxy itself is uncertain?* A straightforward additive reward, $r_{\text{task}} + \lambda r_{\text{vis}}$, implicitly trusts the visual verifier equally on clear and ambiguous inputs. That is undesirable under distribution shift, weak verifiers, or sparse exact feedback. Our proposal, CAVE-RL, treats the visual reward as evidence rather than ground truth. An ensemble produces a visual posterior and a disagreement signal; predictive entropy captures input ambiguity; the two are combined into a reliability score that gates the visual component. The visual reward is additionally centered relative to a frozen reference policy, preventing a prompt-dependent action prior from masquerading as evidence. A KL anchor limits policy drift.

Our experiments are deliberately designed to expose reward unreliability rather than hide it behind a large pretrained model. We construct five image-conditioned instruction tasks over the public handwritten-digits corpus and hold out one natural-language paraphrase per task. The policy starts

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.

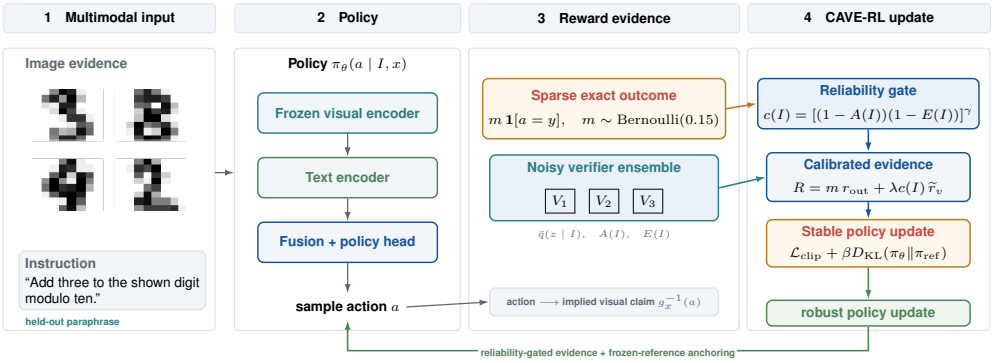


Figure 1: **CAVE-RL.** A multimodal policy receives an image and language instruction, then samples an action. Exact task feedback is observed on only a subset of rollouts. A noisy visual-verifier ensemble supplies dense proxy evidence. Predictive entropy and ensemble disagreement determine how strongly that evidence enters the reward; PPO-style updates remain anchored to a frozen reference policy. The controlled benchmark uses real images and a held-out language paraphrase to separately stress visual and linguistic robustness.

from supervised vision pretraining and a small multimodal cold start. During RL, only 15% of rollouts receive exact task outcomes, while three visual verifiers are trained with 60% corrupted labels. We evaluate clean accuracy, Gaussian visual noise, central occlusion, held-out prompts, a severe combined shift, negative log-likelihood, expected calibration error (ECE), and a proxy-oracle gap.

The resulting picture is consistent across five seeds. Relative to SFT, CAVE-RL gains +0.69 pp clean accuracy, +4.39 pp under occlusion, and +3.88 pp under severe shift, with a -4.16 pp reduction in ECE. Against PPO with the same 15% exact-feedback budget, it gains +0.70 pp in the five-condition robust average. Dense-outcome PPO remains stronger, but CAVE-RL recovers 74% of its robustness improvement over SFT while using 85% fewer exact outcome labels during RL. Across the five-seed study, the largest gains consistently appear under visual corruption, while held-out prompt performance remains comparatively stable, indicating that the improvement is concentrated in the visual-reward pathway rather than language memorization.

Our contributions are threefold. **(1)** We formulate confidence-adaptive visual evidence as a general reward interface for multimodal RL. **(2)** We derive a reliability gate from entropy and ensemble disagreement and combine it with reference-centered visual rewards and KL-stable policy updates. **(3)** We execute a five-seed controlled experiment with sparse exact rewards, deliberately corrupted visual verifiers, held-out language prompts, and visual distribution shifts, and analyze robustness, calibration, reward efficiency, and proxy over-optimization.

2 Related Work

RL and preference optimization. PPO is a standard policy-gradient method for RLHF (Schulman et al., 2017; Ouyang et al., 2022); DPO replaces online RL with a direct preference objective (Rafailov et al., 2023). GRPO removes a learned value critic by using group-relative advantages and has become common in reasoning-centric post-training (Shao et al., 2024; DeepSeek-AI, 2025). AI feedback can further reduce reliance on human preference labels (Bai et al., 2022; Lee et al., 2024). Our setting differs in that the auxiliary reward is explicitly visual and known to be noisy.

Multimodal alignment and reward models. RLHF-V uses correctional human feedback for trustworthy multimodal alignment (Yu et al., 2024); RLAIF-V replaces human preference collection with AI feedback (Yu et al., 2025); VLFeedback provides large-scale multimodal preference data (Li et al., 2024); and MM-RLHF scales human preference supervision (Zhang et al., 2025). Visual-PRM400K studies process reward modeling over visual reasoning traces (Wang et al., 2026a), while R1-Reward directly optimizes a multimodal reward model with RL (Zhang et al., 2026b). Deep ensembles provide a simple, effective uncertainty estimator (Lakshminarayanan et al., 2017), and

modern calibration work shows why confidence under shift should not be taken for granted (Guo et al., 2017; Ovadia et al., 2019). CAVE-RL uses these uncertainty ideas at the reward-composition layer.

Reasoning-centric multimodal RL. Vision-R1 combines cold-start data and RL (Huang et al., 2026); cold-start ablations show that SFT and RL effects interact strongly (Wei et al., 2025); ReVisual-R1 further studies staged RL and gradient stagnation (Chen et al., 2026). Perception-R1 adds an explicit visual-perception reward (Xiao et al., 2026); PAPO changes the policy objective to increase visual sensitivity (Wang et al., 2026b); RewardMap targets sparse rewards with dense visual supervision (Feng et al., 2026); DeepEyes learns active visual inspection (Zheng et al., 2026); STVG-R1 uses task-driven spatial-temporal grounding rewards (Zhang et al., 2026a); and GuirlVG emphasizes stabilization against reward over-optimization (Kang et al., 2026). These works motivate our focus on reward reliability rather than only reward availability. Evaluation practice commonly uses MathVista, MMMU, and MMStar, with toolkits such as VLMEvalKit standardizing many MLLM comparisons (Lu et al., 2024; Yue et al., 2024; Chen et al., 2024; Duan et al., 2024). Widely used open backbones include Qwen2.5-VL and LLaVA-family models (Bai et al., 2025; Liu et al., 2023).

3 METHOD

3.1 PROBLEM SETUP

Let I be an image, x a language instruction, and a a discrete or structured policy action sampled from $\pi_\theta(a \mid I, x)$. An exact task reward $r_{\text{out}}(a, y) \in [0, 1]$ exists but is observed only when $m \sim \text{Bernoulli}(\rho)$ equals one. Our experiments use $\rho = 0.15$. We additionally have K imperfect visual verifiers $q_k(z \mid I)$ over a visual latent state $z \in \{1, \dots, C\}$. The instruction defines a mapping from action to the visual claim it implies,

$$z = g_x^{-1}(a). \tag{1}$$

For a general MLLM, z can be an object identity, coordinate bin, attribute, intermediate answer, or another checkable visual claim.

3.2 VISUAL EVIDENCE AND UNCERTAINTY

The ensemble posterior and per-class standard deviation are

$$\bar{q}(z \mid I) = \frac{1}{K} \sum_{k=1}^{K} q_k(z \mid I), \qquad \sigma(z \mid I) = \sqrt{\frac{1}{K} \sum_{k=1}^{K} \left(q_k(z \mid I) - \bar{q}(z \mid I)\right)^2}. \tag{2}$$

We decompose uncertainty into a normalized predictive-entropy term $A(I)$ and an ensemble-disagreement term $E(I)$:

$$A(I) = \frac{H[\bar{q}(\cdot \mid I)]}{\log C}, \tag{3}$$

$$E(I) = \frac{H[\bar{q}(\cdot \mid I)] - \frac{1}{K} \sum_k H[q_k(\cdot \mid I)]}{\log C}. \tag{4}$$

The second expression is the Jensen–Shannon disagreement induced by the ensemble. We define visual-reward reliability as

$$c(I) = \left[(1 - A(I))(1 - E(I))\right]^\gamma, \qquad c(I) \in [0, 1]. \tag{5}$$

A low-entropy but mutually inconsistent ensemble is down-weighted, as is a consensually uncertain ensemble. This distinction matters under shift, where epistemic and aleatoric uncertainty can behave differently (Ovadia et al., 2019).

The raw visual reward for action a is $r_v(a, I, x) = \bar{q}(g_x^{-1}(a) \mid I)$. Directly optimizing it can amplify action priors. We therefore center it with respect to a frozen reference policy π_{ref}:

$$\widetilde{r}_v(a, I, x) = r_v(a, I, x) - \mathbb{E}_{a' \sim \pi_{\text{ref}}(\cdot \mid I, x)}[r_v(a', I, x)]. \tag{6}$$

The CAVE-RL reward is

$$R(a, I, x) = m\, r_{\text{out}}(a, y) + \lambda c(I) \widetilde{r}_v(a, I, x). \tag{7}$$

Algorithm 1 Confidence-Adaptive Visual Evidence RL

Require: reference policy π_{ref}, verifiers $\{q_k\}_{k=1}^{K}$, exact-reward rate ρ
1: **for** each RL minibatch (I, x, y) **do**
2: sample $a \sim \pi_\theta(\cdot \mid I, x)$ and $m \sim \text{Bernoulli}(\rho)$
3: compute $\bar{q}$, entropy $A(I)$, disagreement $E(I)$, and $c(I)$ by Eq. 5
4: map a to implied visual claim $g_x^{-1}(a)$ and compute $\widetilde{r}_v$ by Eq. 6
5: form $R = mr_{\text{out}} + \lambda c(I)\widetilde{r}_v$
6: update π_θ with Eq. 9
7: **end for**

The first term remains authoritative whenever exact feedback is available; the second supplies dense visual evidence in proportion to estimated reliability.

A conservative lower confidence score $[\bar{q} - \tau\sigma]_+$ is also possible. Under a sub-Gaussian approximation for verifier error, the probability that a one-sided confidence bound is overly optimistic decreases as $\exp(-\tau^2/2)$. We evaluate this option on validation data, but the selected configuration uses $\tau = 0$: explicit entropy/disagreement gating was sufficient in our controlled setting and avoided discarding useful evidence.

3.3 Stable policy update

We use a clipped PPO-style objective with a reference KL penalty,

$$\mathcal{L}_{\text{clip}} = -\mathbb{E}\left[\min\left(r_t\widehat{A}_t, \text{clip}(r_t, 1-\epsilon, 1+\epsilon)\widehat{A}_t\right)\right], \tag{8}$$

$$\mathcal{L} = \mathcal{L}_{\text{clip}} + \beta D_{\text{KL}}(\pi_\theta \| \pi_{\text{ref}}) - \eta H(\pi_\theta), \tag{9}$$

where $r_t = \pi_\theta(a_t \mid I_t, x_t)/\pi_{\text{ref}}(a_t \mid I_t, x_t)$ and $\widehat{A}_t$ is the normalized batch advantage computed from Eq. 7. Reference anchoring is motivated by the broader reward-overoptimization literature (Gao et al., 2023) and by multimodal evidence that aggressive optimization can exploit weak verifiers (Yao et al., 2026; Kang et al., 2026).

4 Experimental Setup

4.1 Controlled multimodal post-training benchmark

We use the public handwritten digits dataset distributed with scikit-learn (Pedregosa et al., 2011): 1,797 grayscale images of real handwritten digits. The fixed image split contains 1,077 train, 360 validation, and 360 test images. We expand these source images into a language-conditioned multimodal benchmark by pairing each image with five transformations—*identify*, *nine-complement*, *next digit modulo 10*, *previous digit modulo 10*, and *add three modulo 10*—and four natural-language paraphrases per task. The first three paraphrases are available during training, while the fourth is held out from training. This produces 16,155 training contexts, 7,200 validation contexts, and 7,200 test contexts (30,555 image-instruction instances in total), while preserving disjoint source-image splits. Thus, success requires both visual recognition and instruction-conditioned transformation rather than memorizing repeated labels.

The policy contains a two-layer CNN visual encoder, a token-embedding text encoder, and a fusion MLP policy head. The visual encoder is first trained for digit recognition, then frozen. A multimodal cold start trains the text/fusion head on 45% of training images and three prompt variants, corresponding to 7,200 supervised multimodal contexts. RL then samples from the full 16,155-context training pool. Each reported test condition evaluates all 7,200 test image-instruction contexts per seed and method.

Noisy visual verifiers. We train three independently initialized CNN verifiers. During verifier training, 60% of digit labels are replaced by uniformly random labels. During RL, 40% of minibatch images receive a visual perturbation; alternating updates additionally attenuate the central 3×3 region. These choices create a reward source that is useful but materially unreliable.

Table 1: **Five-seed results** (mean ± standard deviation, %). Exact feedback is the fraction of RL rollouts exposing the true task outcome. Robust Avg. averages clean, Gaussian noise, occlusion, held-out prompt, and severe shift. Lower ECE is better.

Method	Exact feedback	Clean	Gaussian	Occlusion	Held-out prompt	Severe shift	Robust Avg.	ECE ↓
SFT	–	75.86 ± 2.06	69.83 ± 1.50	34.46 ± 2.05	36.52 ± 6.67	32.21 ± 2.46	49.78 ± 1.68	12.49 ± 1.86
PPO-Sparse	15%	76.06 ± 1.95	70.53 ± 1.45	37.78 ± 2.03	36.83 ± 6.37	34.99 ± 2.71	51.24 ± 2.21	10.28 ± 1.50
Grounded-PPO	15%	76.47 ± 2.07	71.05 ± 1.66	38.78 ± 1.75	36.61 ± 7.19	36.32 ± 2.20	51.84 ± 2.12	8.20 ± 1.39
Grounded-Ens	15%	76.43 ± 2.08	71.28 ± 1.65	38.76 ± 1.79	36.66 ± 6.95	35.91 ± 1.89	51.80 ± 2.31	$\mathbf{8.11 \pm 1.55}$
CAVE-RL	**15%**	$\mathbf{76.55 \pm 2.11}$	$\mathbf{71.32 \pm 1.62}$	$\mathbf{38.85 \pm 1.72}$	$\mathbf{36.89 \pm 6.85}$	$\mathbf{36.09 \pm 1.76}$	$\mathbf{51.94 \pm 2.24}$	8.34 ± 1.59
PPO-Full (oracle reward)	100%	77.04 ± 1.77	71.59 ± 1.57	40.90 ± 1.12	36.68 ± 7.08	37.31 ± 1.53	52.71 ± 2.04	8.88 ± 1.41

RL configuration. All RL methods start from the same cold-start policy and run for 150 updates with batch size 112. Exact outcome reward is observed with probability $\rho = 0.15$ except for PPO-Full, which receives it on 100% of rollouts and serves as an oracle-feedback upper bound. We use $\lambda = 6$, PPO clip range $[0.8, 1.2]$, $\beta = 0.03$, and entropy coefficient 0.002. The confidence exponent is selected using validation-only robust accuracy,

$$\gamma^\star = \arg \max_{\gamma \in \{0.25, 0.5, 0.75, 1.0\}} \mathrm{RobustAvg}_{\mathrm{val}}(\gamma), \tag{10}$$

which yields $\gamma^\star = 0.25$; this value is then fixed before the five-seed test runs.

Baselines. SFT is the cold-start policy. PPO-Sparse uses only the 15% observed exact outcomes. Grounded-PPO adds a dense visual reward from one noisy verifier. Grounded-Ens replaces it with the ensemble mean but does not reliability-gate the reward. CAVE-RL adds Eq. 5 on top of the ensemble-centered reward. PPO-Full uses dense exact outcomes and no proxy reward.

Metrics. We report clean accuracy, Gaussian-noise accuracy, central-occlusion accuracy, held-out-prompt accuracy, severe-shift accuracy, and their unweighted mean (*Robust Avg.*). We additionally report ECE (Guo et al., 2017), NLL, and a proxy-oracle gap under severe shift,

$$G = \Pr\left[g_x^{-1}(a) = \arg\max_z \bar{q}(z \mid I)\right] - \Pr[a = y], \tag{11}$$

which is high when actions agree with the visual proxy more often than with the true task target.

5 Results

5.1 Main comparison

Table 1 separates reward-budget efficiency from absolute performance. Dense exact-feedback PPO remains strongest overall, as expected. The relevant comparison for CAVE-RL is the shared 15% exact-feedback budget. Against PPO-Sparse, CAVE-RL gains 0.49 pp clean, 0.79 pp Gaussian, 1.07 pp occlusion, 1.10 pp severe-shift, and 0.70 pp in Robust Avg. Against SFT, the Robust Avg. improvement is 2.17 pp. The oracle PPO gain over SFT is 2.93 pp. Defining reward-budget recovery as

$$\eta_{\mathrm{rec}} = \frac{R_{\text{CAVE-RL}} - R_{\mathrm{SFT}}}{R_{\text{PPO-Full}} - R_{\mathrm{SFT}}}, \tag{12}$$

we obtain $\eta_{\mathrm{rec}} = 74\%$: most of the oracle robustness improvement is recovered while exact outcomes are observed on only 15% of RL rollouts.

The largest gains appear when vision is corrupted. Under central occlusion, CAVE-RL improves SFT by 4.39 pp; under severe shift, by 3.88 pp. Held-out-prompt accuracy changes only modestly, indicating that the method primarily improves the visual-reward pathway rather than language generalization. Calibration improves substantially: ECE falls from 12.49% (SFT) and 10.28% (sparse PPO) to 8.34%.

5.2 Where does the gain come from?

Figure 3 plots improvement over the shared SFT initialization. PPO-Sparse already benefits from sparse exact feedback, especially on occlusion and severe shift. Adding visual evidence gives a

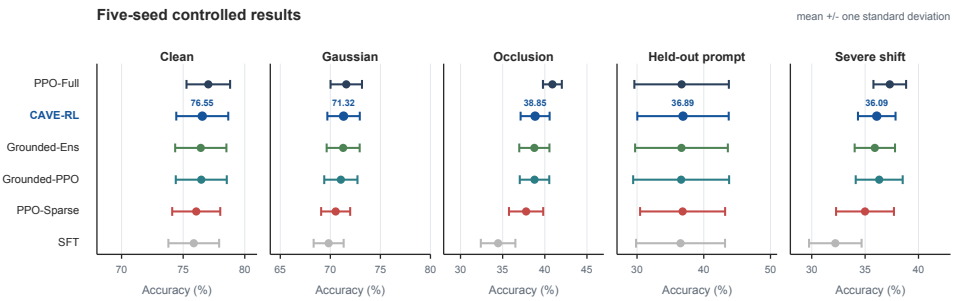


Figure 2: **Main controlled results across five seeds.** Points are means and error bars are one standard deviation. The oracle-feedback PPO baseline observes exact outcome reward on every rollout; all grounded methods and PPO-Sparse observe it on only 15%.

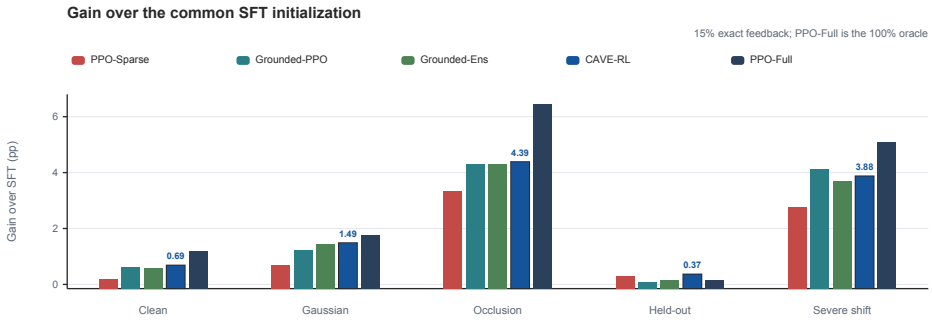


Figure 3: **Gain over the common SFT initialization.** Reliability-aware visual rewards close part of the gap to dense oracle-reward PPO while using 85% fewer exact outcome observations during RL.

further step. Ensemble averaging and confidence adaptation are smaller, second-order changes, but they improve robustness without requiring additional exact outcome labels. The key result is therefore not that uncertainty gating creates a large clean-accuracy jump; it makes a dense visual proxy usable under a severely limited exact-reward budget.

A useful warning emerges from Table 2. Visual grounding can increase proxy agreement faster than oracle accuracy, so a grounded reward should not be judged by its own score. Relative to Grounded-Ens, confidence gating lowers the severe-shift proxy-oracle gap by 0.49 pp and improves Robust Avg. by 0.14 pp. These are modest five-seed effects, so we treat them as mechanism evidence rather than a claim of universal superiority. The stronger and more stable conclusion is the gap between sparse-outcome PPO and visually grounded methods under the same exact-feedback budget.

5.3 Task coverage and training behavior

The five instruction transforms are not equally difficult, but improvements are distributed rather than concentrated in one task. Mean test accuracies for CAVE-RL are 70.61 (identify), 83.64 (nine-complement), 82.04 (next), 75.71 (previous), and 70.76 (shift-by-three). PPO-Sparse is 70.81, 83.75, 80.26, 75.15, and 70.35 respectively. Thus the largest CAVE-RL gain occurs on the "next" transformation, while identify accuracy is essentially unchanged.

Figure 4 shows on-policy sampled correctness and KL drift during RL. The KL term remains small across methods, which is important because weak proxy rewards become more dangerous as the policy moves outside the verifier's training distribution. This observation is consistent with the broader reward-overoptimization literature (Gao et al., 2023) and recent MLLM findings on stabilization (Kang et al., 2026).

Table 2: **Component comparison under the 15% exact-feedback budget.** The proxy-oracle gap is measured under severe visual shift; lower is better. CAVE-RL slightly reduces the ensemble proxy gap while preserving its accuracy gains.

Method	Robust Avg.	ECE ↓	NLL ↓	Proxy-oracle gap ↓
PPO-Sparse	51.24 ± 2.21	10.28 ± 1.50	0.843 ± 0.029	7.01 ± 4.25
Grounded-PPO	51.84 ± 2.12	8.20 ± 1.39	0.828 ± 0.035	9.07 ± 3.15
Grounded-Ens	51.80 ± 2.31	$\mathbf{8.11 \pm 1.55}$	0.823 ± 0.036	9.75 ± 3.96
CAVE-RL	$\mathbf{51.94 \pm 2.24}$	8.34 ± 1.59	$\mathbf{0.820 \pm 0.036}$	9.26 ± 3.95

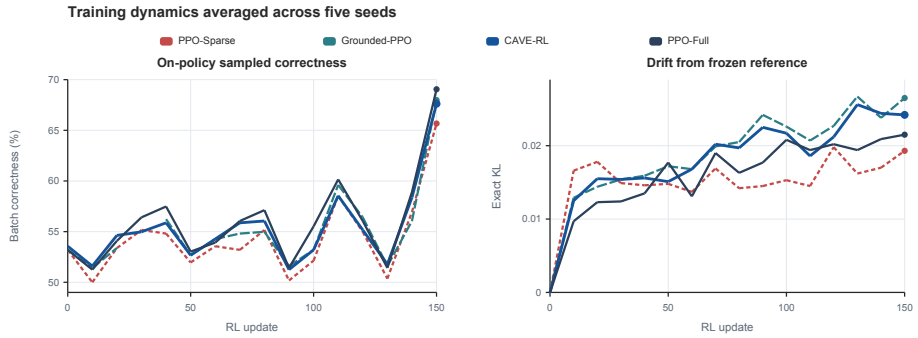


Figure 4: **Training dynamics averaged over five seeds.** Left: sampled task correctness during RL. Right: exact KL divergence from the frozen reference policy.

6 Discussion

Why reliability gating instead of simply averaging verifiers? Ensembling reduces variance, but its mean can still be confidently wrong under shared distribution shift. Equation 5 uses both entropy and disagreement: low entropy alone is insufficient, and agreement alone is insufficient if all ensemble members are uncertain. The five-seed comparison suggests that this additional gating has a small positive robustness effect relative to a raw ensemble. More importantly, it provides a principled interface for stronger multimodal verifiers whose confidence varies substantially by image and reasoning step.

What transfers to large MLLMs? The action-to-visual-claim map in Eq. 1 is simple in our benchmark, but analogous mappings already appear in MLLM systems: object IDs in STVG-R1, GUI coordinates in GuirlVG, visual annotations in Perception-R1, and intermediate verification in multimodal reward-model work (Zhang et al., 2026a; Kang et al., 2026; Xiao et al., 2026; Zhang et al., 2026b). CAVE-RL does not require the visual claim to be a class label; it only requires a verifier distribution or score whose reliability can be estimated.

Interpretation of the oracle baseline. PPO-Full is intentionally advantaged: it sees the true task reward on every rollout. It is not a resource-matched baseline and should not be read as a method that CAVE-RL must beat. Its role is to quantify the maximum benefit available from dense exact feedback in this controlled environment. Recovering 74% of its robustness gain at a 15% exact-feedback rate is the more relevant result.

7 Limitations

The reliability estimate in CAVE-RL is computed at the image level, whereas open-ended multimodal reasoning can contain step-specific uncertainty that changes throughout a trajectory. Extending the gate to token-, region-, or reasoning-step-level evidence is therefore a natural next direction. The verifier ensemble also shares architecture and training data, so correlated errors can reduce the

informativeness of disagreement. Finally, the current implementation uses fixed reward and KL coefficients across tasks; adaptive schedules that react to verifier calibration or policy drift may further improve optimization efficiency.

8 Conclusion

We introduced CAVE-RL, a reliability-aware reward construction for multimodal RL under sparse exact feedback. The method gates dense visual proxy rewards using predictive entropy and ensemble disagreement, centers them against a frozen reference policy, and uses KL-stable PPO updates. Across five random seeds and 30,555 multimodal image-instruction instances, CAVE-RL improves robustness and calibration over SFT and sparse-reward PPO while recovering most of the robustness gain of dense oracle-feedback PPO with only 15% exact outcome feedback. The results further show that proxy agreement and true task accuracy can diverge under visual shift, making uncertainty-aware reward control an important component of multimodal policy optimization.

AI Use Statement

Generative AI tools were used solely for language polishing.

References

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. *arXiv preprint arXiv:2502.13923*, 2025.

Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, et al. Constitutional ai: Harmlessness from ai feedback. *arXiv preprint arXiv:2212.08073*, 2022.

Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Jiaqi Wang, Yu Qiao, Dahua Lin, and Feng Zhao. Are we on the right way for evaluating large vision-language models? In *Advances in Neural Information Processing Systems*, 2024.

Shuang Chen, Hangyu Guo, Zhaochen Su, Yafu Li, Jiacheng Chen, Yulun Wu, Weijie Wang, Zhiyuan Feng, Xiaoye Qu, and Yu Cheng. Revisual-r1: Advancing multimodal reasoning from optimized cold start to staged reinforcement learning. In *International Conference on Learning Representations*, 2026.

DeepSeek-AI. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

Haodong Duan, Junming Yang, Yuxuan Qiao, Xinyu Fang, Lin Chen, Yuan Liu, Xiaoyi Dong, Yuhang Zang, Pan Zhang, Jiaqi Wang, Dahua Lin, and Kai Chen. Vlmevalkit: An open-source toolkit for evaluating large multi-modality models. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 11198–11201, 2024.

Sicheng Feng, Kaiwen Tuo, Song Wang, Lingdong Kong, Jianke Zhu, and Huan Wang. Rewardmap: Tackling sparse rewards in fine-grained visual reasoning via multi-stage reinforcement learning. In *International Conference on Learning Representations*, 2026.

Leo Gao, John Schulman, and Jacob Hilton. Scaling laws for reward model overoptimization. In *Proceedings of the 40th International Conference on Machine Learning*, 2023.

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*, 2017.

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. In *International Conference on Learning Representations*, 2026.

Weitai Kang, Bin Lei, Gaowen Liu, Caiwen Ding, and Yan Yan. Guirlvg: Incentivize gui visual grounding via empirical exploration on reinforcement learning. In *International Conference on Learning Representations*, 2026.

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. In *Advances in Neural Information Processing Systems*, 2017.

Harrison Lee, Samrat Phatale, Hassan Mansoor, Thomas Mesnard, Johan Ferret, Kellie Ren Lu, Colton Bishop, Ethan Hall, Victor Carbune, Abhinav Rastogi, and Sushant Prakash. Rlaif vs. rlhf: Scaling reinforcement learning from human feedback with ai feedback. In *Proceedings of the 41st International Conference on Machine Learning*, pages 26874–26901, 2024.

Lei Li, Zhihui Xie, Mukai Li, Shunian Chen, Peiyi Wang, Liang Chen, Yazheng Yang, Benyou Wang, Lingpeng Kong, and Qi Liu. Vlfeedback: A large-scale ai feedback dataset for large vision-language models alignment. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 6227–6246, 2024.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *Advances in Neural Information Processing Systems*, 2023.

Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. In *International Conference on Learning Representations*, 2024.

Damiano Marsili and Georgia Gkioxari. No labels, no problem: Training visual reasoners with multimodal verifiers. In *International Conference on Learning Representations*, 2026.

Long Ouyang, Jeff 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*, 2022.

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*, 2019.

Fabian Pedregosa, Gael Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. Scikit-learn: Machine learning in python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. In *Advances in Neural Information Processing Systems*, 2023.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. In *arXiv preprint arXiv:1707.06347*, 2017.

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

Weiyun Wang, Zhangwei Gao, Lianjie Chen, Zhe Chen, Jinguo Zhu, Xiangyu Zhao, Yangzhou Liu, Yue Cao, Shenglong Ye, Xizhou Zhu, Lewei Lu, Haodong Duan, Yu Qiao, Jifeng Dai, and Wenhai Wang. Visualprm400k: An effective dataset for training multimodal process reward models. In *International Conference on Learning Representations*, 2026a.

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

Lai Wei, Yuting Li, Kaipeng Zheng, Chen Wang, Yue Wang, Linghe Kong, Lichao Sun, and Weiran Huang. Advancing multimodal reasoning via reinforcement learning with cold start. *arXiv preprint arXiv:2505.22334*, 2025.

Tong Xiao, Xin Xu, Zhenya Huang, Hongyu Gao, Quan Liu, Qi Liu, and Enhong Chen. Perception-r1: Advancing multimodal reasoning capabilities of mllms via visual perception reward. In *International Conference on Learning Representations*, 2026.

Jiayu Yao, Yiwei Wang, Anmeng Zhang, Zhe Sun, Songsong Wang, Lingrui Mei, Yuyao Ge, and Shenghua Liu. Multimodal reward hacking in reinforcement learning. *arXiv preprint arXiv:2607.09492*, 2026.

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

Tianyu Yu, Haoye Zhang, Qiming Li, Qixin Xu, Yuan Yao, Da Chen, Xiaoman Lu, Ganqu Cui, Yunkai Dang, Taiwen He, Xiaocheng Feng, Jun Song, Bo Zheng, Zhiyuan Liu, Tat-Seng Chua, and Maosong Sun. Rlaif-v: Open-source ai feedback leads to super gpt-4v trustworthiness. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19985–19995, 2025.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, Cong Wei, Botao Yu, Ruibin Yuan, Renliang Sun, Ming Yin, Boyuan Zheng, Zhenzhu Yang, Yibo Liu, Wenhao Huang, Huan Sun, Yu Su, and Wenhu Chen. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9556–9567, 2024.

Xiaowen Zhang, Zhi Gao, Licheng Jiao, Lingling Li, and Qing Li. Stvg-r1: Incentivizing instance-level reasoning and grounding in videos via reinforcement learning. In *International Conference on Learning Representations*, 2026a.

Yifan Zhang, Tao Yu, Haochen Tian, Chaoyou Fu, Peiyan Li, Jianshu Zeng, Wulin Xie, Yang Shi, Huanyu Zhang, Junkang Wu, Xue Wang, Yibo Hu, Bin Wen, Tingting Gao, Zhang Zhang, Fan Yang, Di Zhang, Liang Wang, and Rong Jin. Mm-rlhf: The next step forward in multimodal llm alignment. In *Proceedings of the 42nd International Conference on Machine Learning*, pages 76625–76654, 2025.

YiFan Zhang, Xingyu Lu, Xiao Hu, Chaoyou Fu, Bin Wen, Tianke Zhang, Changyi Liu, Kaiyu Jiang, Kaibing Chen, Kaiyu Tang, Haojie Ding, Jiankang Chen, Fan Yang, Tingting Gao, Di Zhang, Guorui Zhou, and Liang Wang. R1-reward: Training multimodal reward model through stable reinforcement learning. In *International Conference on Learning Representations*, 2026b.

Ziwei Zheng, Minghao Yang, Jack Hong, Chenxiao Zhao, Guohai Xu, Le Yang, Chao Shen, and XingYu. Deepeyes: Incentivizing thinking with images via reinforcement learning. In *International Conference on Learning Representations*, 2026.

A Implementation and Reproducibility Details

Data splits and prompts. The split seeds are fixed in code. The train/validation/test source-image counts are 1,077/360/360. Four prompt paraphrases are defined for each of the five transforms; variants 0–2 are used for training construction, and variant 3 is reserved from training. The resulting context pools contain 16,155 training instances, 7,200 validation instances, and 7,200 test instances. Each test image is evaluated with all five tasks and all four paraphrases for the main test set; the held-out metric filters to variant 3 only.

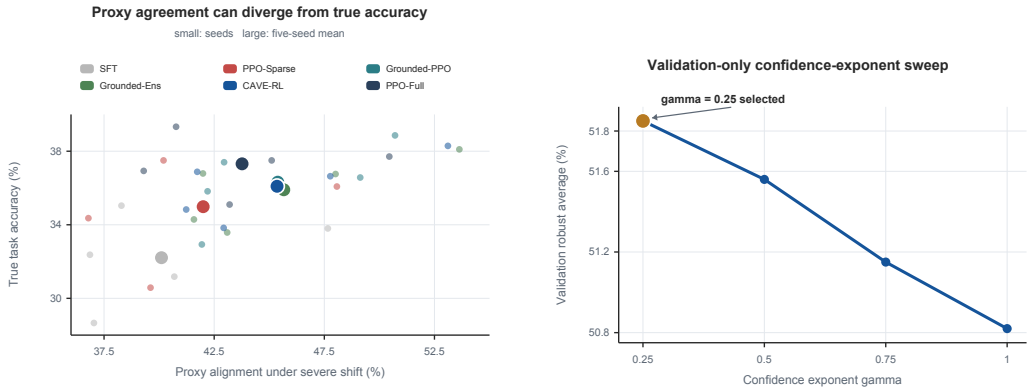
Optimization. The visual encoder pretraining uses Adam with learning rate 2×10^{-3} for 14 epochs. The multimodal cold start trains the non-frozen policy parameters for 9 epochs at 1.2×10^{-3}. RL uses Adam at 5×10^{-4} for 150 updates, batch size 112, gradient norm clipping at 1.0, PPO ratio clipping to $[0.8, 1.2]$, $\beta = 0.03$, and entropy coefficient 0.002. Three visual verifiers are trained independently for 9 epochs with 60% random label corruption.

B Additional Task-Wise Results

Table 3: Mean task accuracy across five seeds (%).

Method	Identify	9-complement	Next	Previous	Shift+3
SFT	71.24	83.43	80.49	73.46	70.68
PPO-Sparse	70.81	83.75	80.26	75.15	70.35
Grounded-PPO	70.58	83.56	82.53	75.49	70.18
Grounded-Ens	70.53	83.40	82.01	75.58	70.62
CAVE-RL	70.61	83.64	82.04	75.71	70.76
PPO-Full	71.24	84.14	82.62	76.11	71.11

C Additional Diagnostics



(a) Proxy alignment versus true severe-shift accuracy. (b) Validation-only confidence-exponent sweep.

Figure 5: **Additional diagnostics.** Left: agreement with an imperfect visual proxy need not track oracle task accuracy, motivating explicit reliability control. Right: the validation sweep selects $\gamma = 0.25$ before final test evaluation.