CONTEXTRL: ADDRESSING POLICY INFORMATION BOTTLENECKS WITH CONTEXT-AUGMENTED RL

Anonymous authors
Paper under double-blind review

ABSTRACT

Reinforcement learning with verifiable rewards (RLVR) has become an important route for improving multimodal large language models (MLLMs), using reward signals provided by verifiers as training supervision. Its effectiveness depends on both reliable rewards and positive samples. In multimodal tasks, final-answer agreement can reward flawed reasoning, while difficult queries may yield no correct response within the sampling budget. Our analysis identifies two coupled information bottlenecks that restrict the quality and availability of policy supervision. To address these bottlenecks, we propose **ContextRL**, a framework that combines context-augmented reward modeling and context-augmented sampling. Context-augmented reward modeling uses full reference solutions to check visual grounding and reasoning, reducing false-positive supervision and generating mistake reports. Context-augmented sampling uses mistake reports to guide further attempts on all-negative groups. Training Qwen3-VL-8B with ContextRL, we achieve average gains of 2.58% and 1.09% over GRPO and DAPO, respectively, across 11 perception and reasoning benchmarks, and surpass Qwen3-VL-32B in perception. Our in-depth analysis shows that contextual information improves reward accuracy and the availability of positive samples, while revealing how false-positive rewards undermine policy learning. These findings support context augmentation for improving RLVR supervision.

1 INTRODUCTION

Multimodal large language models (MLLMs) connect visual representations with language generation, supporting tasks from image description to visual question answering. Vision–language pretraining and visual instruction tuning have broadened these capabilities (Alayrac et al., 2022; Liu et al., 2023; Bai et al., 2023; 2025). Recent work increasingly uses reinforcement learning to strengthen multimodal reasoning and interaction with visual tools (Huang et al., 2025; Hong et al., 2025; Zhang et al., 2026b), extending post-training beyond imitation of demonstration responses.

Reinforcement learning with verifiable rewards (RLVR) improves a policy through an iterative cycle of sampling, reward computation, and policy optimization (Shao et al., 2024; Guo et al., 2025). First, the policy generates a group of candidate responses for each query. A verifier then evaluates their correctness and assigns reward signals. Finally, methods such as GRPO and DAPO convert reward differences within each group into advantages that guide policy updates, encouraging higher-reward responses and discouraging lower-reward ones (Shao et al., 2024; Yu et al., 2025). The policy thus learns from the responses it generates and the distinctions made by the verifier. In multimodal tasks, these distinctions must reflect visual grounding as well as reasoning quality (Zhang et al., 2026a). This workflow exposes two potential bottlenecks: whether positive rewards reliably identify valid responses, and whether sampling produces correct responses to reinforce.

Bottleneck I: reward reliability. A correct final answer can coexist with flawed reasoning or hallucinated visual evidence, causing final-answer verification to admit false positives. Our logged counting and reasoning cases illustrate how such rewards reinforce invalid responses. We define *reward reliability* as the agreement between verifier rewards and true response correctness, including reasoning and visual grounding; unreliable rewards limit supervision quality.

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.

Bottleneck II: sampling reachability. Even with a reliable verifier, the policy cannot improve when difficult queries yield no positive samples, because all-negative groups provide no reward contrast. We define *sampling reachability* as the probability that the current policy samples at least one correct response for a query within a fixed budget; low reachability restricts positive supervision. More attempts increase the budget, while feedback can change the sampling distribution.

These bottlenecks interact. When correct responses are rare, even a modest false-positive rate can contaminate much of the rewarded pool. Removing a group's sole false positive may reveal an all-negative group that still needs recovery. We analyze this coupling to determine where additional context can improve the reliability and availability of policy supervision.

We propose **ContextRL**, which augments context at both operations to overcome these information bottlenecks. **Context-Augmented Reward modeling (CAR)** supplies full reference solutions to check intermediate reasoning and visual grounding, targeting false-positive rewards. **Context-Augmented Sampling (CAS)** uses mistake reports to guide further attempts specifically for all-negative groups. This directs assistance toward queries that ordinary sampling has failed to solve. Finally, *context rollback* pairs accepted, self-contained recoveries with the original query and initial negatives. Retaining negatives restores within-query reward contrast; removing feedback from the training input directs learning toward the original inputs available at deployment.

To evaluate the effectiveness of ContextRL, we train Qwen3-VL-8B-Instruct on visual QA and multimodal mathematics. Across eleven benchmarks, ContextRL reaches perception/reasoning averages of 65.22/62.37, compared with 64.74/60.68 for DAPO. Human annotation of 4,000 responses shows an 8.60% reward-accuracy gain; training logs record recovered positives for 8.56% of all queries. These measure reward reliability and recovered coverage, respectively. CAR/CAS ablations assess whether both interventions contribute to final-policy performance on original inputs.

We summarize our contributions as follows:

- **A coupled analysis of policy information bottlenecks.** We connect reward reliability and sampling reachability, showing why rare correct responses demand stricter false-positive control.
- **Targeted context augmentation for RLVR.** CAR uses reference solutions to improve rewards; CAS uses mistake reports to recover positives from all-negative groups.
- **Evidence connecting supervision to policy performance.** Evaluation on 11 benchmarks and component analyses link reward reliability and sample recovery to policy gains.

2 Policy Information Bottlenecks

The policy learns from sampled responses weighted by verifier rewards. We formalize this process and analyze how reward reliability and sampling reachability constrain policy learning.

2.1 RLVR Formalization

For a multimodal query x and reference annotation t, a behavior policy π_{old} samples G responses $y_k \sim \pi_{\text{old}}(\cdot \mid x)$. A verifier assigns binary rewards $R_k = V(x, y_k, t)$. For a group with nonconstant rewards, GRPO normalizes them to obtain the policy advantages

$$A_k = \frac{R_k - \mu_x}{\sigma_x}, \tag{1}$$

where μ_x and σ_x are the group's reward mean and standard deviation, and uses a clipped policy objective with KL regularization (Shao et al., 2024). Responses y_k and their reward-derived weights jointly determine policy supervision. Appendix A.2 details the workflow.

2.2 Two Information Bottlenecks for the Policy

We use *information bottleneck* operationally: a restriction on the availability or reliability of training information under finite sampling and verification resources. This separates two failure modes without asserting a universal information-theoretic capacity limit for RLVR.

Let $C(x, y) \in \{0, 1\}$ indicate true correctness, including valid reasoning and visual grounding, and define the policy's success probability as $p_x = \Pr_{y \sim \pi_{\text{old}}(\cdot|x)}[C(x, y) = 1]$.

2.2.1 BOTTLENECK I: REWARD RELIABILITY

A correct final answer is insufficient evidence of a correct reasoning trace. Checking an intermediate claim may require visual interpretation, a mathematical derivation, or knowledge that is difficult for a finite verifier to recover from a short reference answer. Fuller solutions can aid this verification.

For a fixed x and policy, define the verifier's false-positive and false-negative rates as $\alpha_x = \Pr(R = 1 \mid C = 0, x)$ and $\beta_x = \Pr(R = 0 \mid C = 1, x)$. The positive-reward rate is

$$u_x = p_x(1 - \beta_x) + (1 - p_x)\alpha_x, \tag{2}$$

and, for $u_x > 0$, the correctness of rewarded responses is

$$\Pr(C = 1 \mid R = 1, x) = \frac{p_x(1 - \beta_x)}{u_x}. \tag{3}$$

The numerator counts correct responses receiving positive rewards; the denominator also includes false positives. A larger rewarded pool can thus reflect better sampling or noisier verification. Equation (3) motivates evaluating positive-label quality alongside overall reward accuracy.

Context and the scope of the analysis. If the final answer a is contained in a full solution s, the ideal observation $T_1 = (x, y, s)$ includes the information in $T_0 = (x, y, a)$. Consequently $H(C \mid T_1) \leq H(C \mid T_0)$. This does not guarantee that a learned verifier uses the additional context correctly. When x and y already determine the truth, the practical bottleneck may be computational or perceptual difficulty, rather than missing information. We therefore evaluate context augmentation through observed judgments and policy outcomes.

2.2.2 BOTTLENECK II: SAMPLING REACHABILITY

Under independent sampling, the probability of obtaining at least one correct response is

$$q_x(G) = 1 - (1 - p_x)^G. \tag{4}$$

For small p_x and a limited G, many groups contain no correct example. Even a perfect verifier cannot select a correct solution absent from the group. With binary group-relative rewards, an all-negative group has no reward-relative contrast; an all-positive group also lacks such contrast. This statement concerns the reward component of the update, not the KL term or every possible RL objective. Verifier errors can also yield nonconstant rewards within groups of incorrect responses.

2.2.3 IMPLICATIONS FOR TARGETED CONTEXT AUGMENTATION

Context utility depends on the policy, verifier, and query difficulty. Under the sampling model, $q_x(G) \leq Gp_x$: a very small success probability limits coverage at a fixed budget. Feedback aims to raise this success probability, with benefits assessed against its generation and use costs.

Reward reliability also becomes more demanding as correct responses become rare. For $0 < p_x < 1$, $\beta_x < 1$, and a desired positive-label purity $0 < \tau < 1$, Equation (3) implies

$$\Pr(C = 1 \mid R = 1, x) \geq \tau \quad \Longleftrightarrow \quad \alpha_x \leq \frac{p_x(1 - \beta_x)(1 - \tau)}{(1 - p_x)\tau}. \tag{5}$$

The relation follows by multiplying by the positive denominator and collecting terms in α_x. For example, if $p_x = 0.1$ and $\beta_x = 0.1$, achieving 90% positive-label purity requires $\alpha_x \leq 1/90 \approx 1.1\%$. This is an illustrative consequence of the model, not an observed error rate. It shows why modest false-positive rates can be consequential on difficult queries. Reference context targets this discrimination problem, while guided sampling seeks to increase access to correct candidates.

Positive rewards do not guarantee within-group contrast. Under independent binary rewards, the probability of a mixed-reward group is $1 - u_x^G - (1 - u_x)^G$ (Appendix A.3.3). Cleaner labels need not monotonically increase this probability, and more contrast need not mean better supervision.

These relations motivate coordinated intervention. Reference context targets false positives; report-guided sampling targets groups without accepted solutions, including those exposed by better verification. Recoveries require reliable judgments, while cleaner judgments alone cannot supply absent successes. Experiments assess reward reliability, recovery coverage, and policy performance; difficulty-dependent benefits and compute efficiency remain empirical questions.

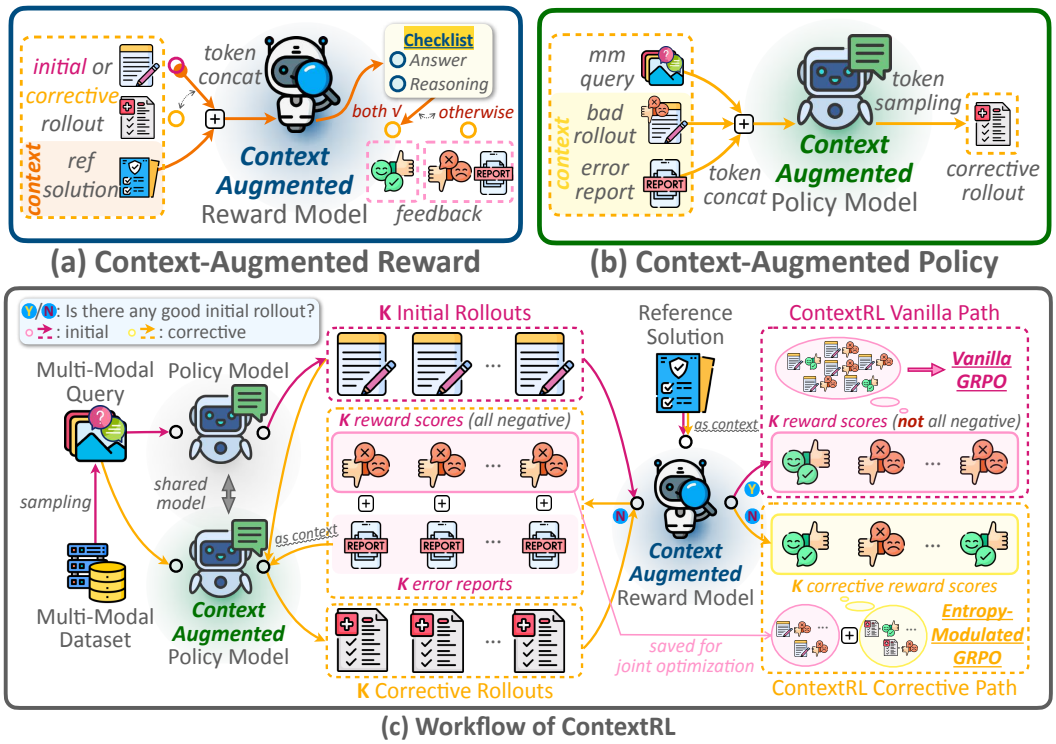


Figure 1: **Overview of ContextRL.** (a) Reference solutions support answer and reasoning checks in CAR. (b) CAS conditions recovery on the query, failed response, and error report. (c) Ordinary and corrective training paths share the policy. Here K denotes group size G; Section 3.3 specifies the corrective update. Final evaluation uses original inputs without reference solutions or reports.

3 CONTEXTRL

Figure 1 outlines ContextRL; Algorithm 1 in Appendix A.1 details its training. ContextRL applies context at each restriction: CAR enriches verification evidence, and CAS changes sampling conditions for all-negative groups. Rollback then trains on accepted recoveries under the original query, connecting assisted data generation to policy performance without assistance at deployment.

3.1 CONTEXT-AUGMENTED REWARD MODEL

Given query x and full solution s (reasoning and final answer), the reward model checks a candidate against the visual input and s, returning a reward and, when applicable, a mistake report:

$$R(x, y) = r_\psi(x, y; s), \qquad M(x, y^-) = h_\psi(x, y^-; s). \tag{6}$$

The instructions ask the model to identify incorrect reasoning without supplying the correct final answer in the report. The reference and report nonetheless provide teacher-derived supervision; their correctness and informational content depend on the evaluator. CAR targets false-positive rewards caused by invalid reasoning or unsupported visual claims, improving the evidence used to weight policy experiences. Appendix D provides the documented instruction templates.

3.2 CONTEXT-AUGMENTED SAMPLING

Stage 1: sample from the original input. The current policy samples $\{y_g^{(1)}\}_{g=1}^{G}$ from $\pi_\theta(\cdot \mid x)$. The context-augmented verifier assigns rewards and reports. If the group contains a positive reward, it supplies the ordinary training group.

Stage 2: recover experiences for all-negative groups. If all first-stage rewards are zero, each failed response y_g^- and its report M_g are appended to the query:

$$y_g^{(2)} \sim \pi_\theta(\cdot \mid x, y_g^-, M_g), \qquad g = 1, \ldots, G. \tag{7}$$

CAS targets groups for which ordinary sampling supplies no accepted solution. Additional ordinary draws increase the number of attempts; the mistake report also changes the conditional distribution of those attempts. We retain verifier-accepted responses that are *self-contained*: they can be read as a solution to x without mentioning the earlier attempt or report. Denote this retained set by

$$\mathcal{P}(x) = \{y_g^{(2)} : R(x, y_g^{(2)}) = 1,\ S(y_g^{(2)}) = 1\}, \tag{8}$$

where S is the standalone-output criterion. This criterion concerns the output's form; it does not imply statistical independence from the feedback used to generate it.

3.3 Using Recovered Experiences to Train the Policy

Ordinary and recovered groups. For a first-stage group containing positive rewards, training uses the original pairs $\mathcal{G}_{\text{on}}(x) = \{(x, y_g^{(1)})\}_{g=1}^G$. For an all-negative group with recovered positives, *context rollback* removes the failed attempt and report from the training input and forms

$$\mathcal{G}_{\text{mix}}(x) = \{(x, y_g^{(1)})\}_{g=1}^G \cup \{(x, y) : y \in \mathcal{P}(x)\}. \tag{9}$$

Retained negatives and recovered positives restore reward-relative contrast for the query. Advantages are computed within this mixed group, using GRPO-style optimization with advantage scaling and selective KL regularization. With $A(x, y)$ denoting the group-relative value, the scaling is

$$\widetilde{A}(x, y) = \lambda A(x, y), \qquad 0 < \lambda < 1, \tag{10}$$

for every response in a mixed group; ordinary groups use unscaled advantages. KL regularization excludes recovered positives. Appendix B gives the optimization settings.

Online updates with context rollback. Both stages sample from the current policy, followed by one update per rollout batch. Recovery uses feedback, whereas rollback trains under the original query. Advantage scaling controls the recovered groups' influence without equating these two conditional distributions. Final evaluation uses original inputs, testing whether contextual assistance improves the policy's unaided performance.

4 Experiments

Our evaluation follows the two-bottleneck analysis. We first measure the final policy on perception and reasoning tasks, then examine false-positive rewards, overall reward reliability, and recovery of all-negative groups. Component ablations assess their effects on original-query performance.

4.1 Experimental Setup

We train Qwen3-VL-8B-Instruct with Qwen3-VL-32B-Instruct (Bai et al., 2025) as the reward model on 29K FineVision (Wiedmann et al., 2025) visual QA and multimodal mathematics examples. We generate full reference solutions during data construction (Appendix B). Baselines are the original 8B, reference SFT, GRPO (Shao et al., 2024), DAPO (Yu et al., 2025), and Qwen3-VL-32B.

The five perception benchmarks are SimpleVQA, MMStar, HalluBench (HallusionBench), HR-8K (HRBench8K), and MME-RW-L (MME-RealWorld-Lite); the six reasoning benchmarks are MVerse (MathVerse), MathVista, LogicVista, We-Math, CharXiv-RQ, and DynaMath (Appendix B.2). Each category Avg in Table 1 weights its ten metric columns equally, rather than weighting benchmarks equally. Gains written with % denote absolute differences between percentage scores. VLMEvalKit uses GPT-4o (Hurst et al., 2024) for original-query evaluation without reference solutions or reports.

We use eight H800 GPUs, ms-swift (Zhao et al., 2025b), global batch size 128, group size eight, a 3M-pixel budget, and maximum response length 16K. SFT runs five epochs at learning rate 10^{-5}; RL runs one at 10^{-6}, with ContextRL/GRPO KL coefficient 0.01. We evaluate the selected SFT and final RL checkpoints. Equal epochs need not match policy tokens, verifier calls, or feedback costs.

Table 1: Perception and reasoning results for the evaluated models. Each Avg. is the unweighted mean of the ten metric columns in its category. Bold and underline indicate the best and second-best distinct scores per column; tied scores share the same rank.

Perception Benchmarks											
Model	**SimpleVQA**	**MMStar**	**HalluBench**			**HR-8K**			**MME-RW-L**		**Avg.**
	overall	*overall*	*aAcc*	*fAcc*	*qAcc*	*single*	*cross*	*overall*	*percept*	*reason*	
Qwen3-VL 8B	44.93	64.06	71.08	53.46	53.19	79.00	61.00	70.00	54.66	41.73	59.31
SFT	46.14	70.87	74.97	56.36	**56.48**	67.75	70.00	71.88	55.95	50.27	62.07
GRPO	46.86	71.87	75.50	58.09	55.60	80.50	70.25	75.38	57.83	50.67	64.25
DAPO	47.62	72.00	75.29	**60.98**	54.73	**82.25**	68.75	75.50	**58.51**	51.73	64.74
ContextRL	48.35	73.00	**75.71**	60.40	**56.48**	80.75	**70.50**	**75.63**	58.17	**53.20**	**65.22**
Qwen3-VL 32B	**53.07**	**74.53**	75.08	55.49	55.38	79.25	**70.50**	74.88	57.14	45.73	64.10

Reasoning Benchmarks											
Model	**MVerse**	**MathVista**	**LogicVista**	**We-Math**		**CharXiv-RQ**			**DynaMath**		**Avg.**
	mini	*mini*	*overall*	*strict*	*loose*	*chart*	*general*	*overall*	*worst*	*overall*	
Qwen3-VL 8B	60.50	76.80	56.82	55.33	72.67	50.36	45.68	46.00	39.52	67.49	57.12
SFT	64.82	76.80	59.06	55.43	73.52	48.38	47.95	47.80	38.32	67.37	57.94
GRPO	64.64	78.80	57.04	61.62	75.90	47.32	43.66	44.90	40.11	67.80	58.17
DAPO	65.17	77.70	60.85	63.81	78.67	51.29	51.13	48.90	41.17	68.08	60.68
ContextRL	69.34	78.90	60.40	**64.48**	**81.24**	51.75	51.23	50.40	47.52	68.42	62.37
Qwen3-VL 32B	**69.40**	**82.40**	**62.20**	63.52	78.00	**55.06**	**58.24**	**55.20**	**49.70**	**76.53**	**65.03**

4.2 Final Policy Performance

Table 1 compares five perception and six reasoning benchmarks. We draw four conclusions.

(1) ContextRL delivers stronger policy improvements. ContextRL ranks first or second on every perception metric and nine of ten reasoning metrics, achieving the best averages among the evaluated 8B models. Its gains over both GRPO and DAPO support addressing the information available to policy learning through verification and sampling. Evaluation uses original queries, showing that contextual assistance during training translates into stronger policy capabilities.

(2) The benefits extend across perception and reasoning. ContextRL improves the base 8B model by 5.91% and 5.25% in perception and reasoning, respectively. SFT, GRPO, and DAPO each show larger gains in perception than reasoning. ContextRL's more balanced improvements support its applicability across both task categories. This pattern is consistent with CAS recovering positives for otherwise unsuccessful queries; the component ablations further examine this mechanism.

(3) Contextual use of references outperforms direct SFT. Reference SFT leaves MathVista unchanged and reduces DynaMath performance, whereas ContextRL improves both and achieves higher category averages. These results favor using references to verify and refine policy-generated responses, allowing the policy to learn from its own responses under reference-informed supervision. Controlled reference-quality ablations would test robustness to imperfect references.

(4) Perception and reasoning retain different model-size gaps. With one quarter of the policy parameters, ContextRL-trained Qwen3-VL-8B surpasses Qwen3-VL-32B in perception by 1.12%. It nearly matches the 32B model on MathVerse and exceeds both We-Math scores, while remaining 2.66% lower on the reasoning average. Thus, contextual supervision substantially narrows the model-size gap, with reasoning remaining the larger challenge. Broader reasoning data would help test how much of this residual gap reflects training coverage.

Table 2: Error-detection rates (%) on the human-annotated false-positive analysis set collected from training logs.

Reward Model	w/o ref	w. answer	w. solution
Qwen3-VL-32B	46.25	50.03	81.98
Qwen3-VL-235B	51.25	54.60	82.12

4.3 Diagnosing False-Positive Rewards

We collect a 30K false-positive analysis set from training logs through model evaluation and human annotation. Responses have correct answers but flawed reasoning or visual grounding and are screened for rejection by both the ContextRL reward model and Qwen3-VL-235B-Instruct. A human audit of 1,000 randomly sampled screened responses confirms errors or hallucinations in 95.8%. We compare no reference (*w/o ref*), final answers (*w. answer*), and full solutions (*w. solution*).

Relative to final-answer context, full-solution context raises error detection from 50.03% to 81.98% for the 32B evaluator and from 54.60% to 82.12% for the 235B evaluator, largely closing their gap. These selected-set rates measure error detection; Section 4.4 evaluates overall reliability on mixed responses. Appendix D provides the evaluation instructions.

4.4 Training-Information Gains: Reliability and Coverage

We assess the two information pathways in Section 2.2: human-annotated correct and incorrect responses measure overall reward reliability, while training logs measure recovery coverage.

Reward reliability. A human-annotated evaluation covers 500 queries with eight responses each, scored under final-answer-based and process-aware rules using Qwen3-VL-32B-Instruct. Of 4,000 responses, 1,793 are fully correct; the rest have an incorrect answer or a flawed process. Process-aware evaluation improves reward accuracy by **8.60%**, from 88.50% to 97.10%, and positive-reward precision from 80.87% to 95.84%. False positives fall from 413 to 76, while false rejections of fully correct responses fall from 47 to 40 (2.62% to 2.23%). Among 414 responses with correct answers but flawed processes, positive judgments fall from 378 to 64. These results support cleaner positive supervision while preserving correct responses, addressing the reliability term in Equation (3). Table 3 reports counts and metrics; Appendix C.2 defines labels and denominators.

Table 3: Reward judgments on 4,000 human-annotated responses from 500 queries. The upper panel counts positive judgments by response class; the lower panel reports reward-classification metrics (%). Precision measures correctness among positive judgments. FPR uses all incorrect responses; FNR uses all fully correct responses.

Reward rule	Incorrect answer	Process error	Fully correct	Total
w. answer	35	378	1,746	2,159
w. solution	12	64	1,753	1,829
Reward rule	**Accuracy**	**Precision**	**FPR**	**FNR**
w. answer	88.50	80.87	18.71	2.62
w. solution	97.10	95.84	3.44	2.23

Recovered training coverage. Training logs record that stage two supplies retained positives for 8.56% of all training queries (Appendix C.3). Each such query previously had an all-negative first-stage group. Mixing its retained positives with the G initial negatives creates nonzero reward-relative advantages for those initial responses. For fixed G and Q queries, the fraction of first-stage responses is

$$\frac{GQ_{\text{recovered}}}{GQ} = \frac{Q_{\text{recovered}}}{Q} = 8.56\%. \tag{11}$$

This measures restored reward contrast, not gradient quality.

Recovery on difficult queries. On 200 difficult queries with all-negative initial groups of eight, both strategies reuse these samples and draw eight more. With CAS, positives cover 133 queries (66.5%), versus 39 (19.5%) with ordinary sampling. At the same 16-rollout allowance, this supports CAS's reachability benefit beyond budget expansion (Appendix C.3.1). These conditional rates differ from the 8.56% coverage of all training queries.

Reward accuracy and recovery coverage use different populations and denominators (Appendix A.3).

4.5 Policy Sensitivity to False-Positive Supervision

To test how supervision errors affect the policy, we vary the fraction of selected false positives from 0% to 30% in 50K-example SFT datasets, mixing them with policy-generated positive responses. The false positives have correct answers but flawed reasoning or visual grounding (Section 4.3).

Table 4: Model performance after SFT with different proportions of false-positive samples in the training data.

Setting	**MME-RW-L**	**MMStar**	**HR-8K**	**MathVista**	**We-Math**
split	*lite*	*overall*	*overall*	*mini*	*strict*
Qwen3-VL 8B	49.61	64.06	70.00	76.80	59.14
all positive	53.93	72.87	73.25	78.60	63.05
w 10% false	53.88	73.93	73.00	78.30	62.76
w 20% false	54.72	73.00	73.50	78.30	61.33
w 30% false	53.41	72.47	73.49	77.00	61.29

Table 4 shows that increasing the false-positive fraction from 0% to 30% lowers MathVista from 78.60 to 77.00 and We-Math from 63.05 to 61.29. MME-RW-L, MMStar, and HR-8K vary non-monotonically, indicating task-dependent sensitivity. The reasoning losses demonstrate why final-answer agreement alone is insufficient for selecting training responses: accepted examples may still teach invalid reasoning. This SFT evidence supports rejecting flawed positive supervision and complements the RL ablations below. Comparisons use the within-table baseline; Appendix C.1 details differences from the main experiment.

4.6 Ablations of the Information Pathways

To link supervision to policy gains, we ablate full-solution reward context (CAR), second-stage sampling (CAS), and mixed-group advantage scaling. The *w/o report* variant uses a 16-rollout ordinary-sampling budget for all-negative queries, matching CAS's two rounds of eight. If positives are found, it forms a training group; otherwise, it trains on all-negative samples.

Removing CAR lowers the eight-metric average from 66.67 to 64.71; removing CAS lowers it to 65.13. Both interventions contribute to aggregate performance. LogicVista and We-Math show substantial losses, while DynaMath improves in both ablations. These differences fit the difficulty- and verifier-dependent view in Section 2.2.3, without validating every prediction.

The 16-rollout ordinary-sampling control reaches 66.35, compared with 66.67 for ContextRL: feedback-guided recovery gives a modest 0.32% final-policy advantage at the matched sampling allowance. Rollout counts do not equalize token or GPU-time costs. Removing advantage scaling lowers the average by 0.43%. Repeated runs would help assess these smaller differences, and training diagnostics would test stability benefits.

Qualitative examples. Logged examples include a correct horse count with an unsupported location (Appendix C.4.2) and a correct option with invalid reasoning (Appendix C.4.1). They illustrate how final-answer agreement can miss flawed experience.

Table 5: Component ablations. CAR: context-augmented reward; CAS: context-augmented sampling. The w/o report row matches CAS's $8 + 8$ allowance with 16 ordinary rollouts; w/o scale removes advantage scaling. Bold and underline mark the top two distinct scores in each column.

Method	SimpleVQA	MMStar	HalluBench	HR-8K	DynaMath	MathVista	LogicVista	We-Math	Avg.
ContextRL	**48.35**	<u>73.00</u>	**64.20**	<u>75.63</u>	68.42	**78.90**	**60.40**	**64.48**	**66.67**
w/o CAR	47.02	72.13	63.13	73.00	<u>69.64</u>	78.60	55.93	58.19	64.71
w/o CAS	46.68	71.27	62.90	74.38	69.58	77.90	55.48	62.86	65.13
w/o report	46.70	**73.07**	63.29	74.63	**70.16**	78.50	<u>60.36</u>	<u>64.10</u>	<u>66.35</u>
w/o scale	<u>47.15</u>	72.80	<u>63.64</u>	**78.25**	68.22	<u>78.80</u>	59.13	61.90	66.24

5 Related Work

5.1 Sampling, Optimization, and Policy Information

Multimodal post-training advances group-relative optimization, response sampling, and visual tool interaction (Huang et al., 2025; Zheng et al., 2025b; Shao et al., 2024; Yu et al., 2025; Zheng et al., 2025a; Wu et al., 2025; Lai et al., 2025; Zhao et al., 2025a), with RL incorporated into broader model pipelines (Kwai Keye Team, 2025; Kimi Team et al., 2025; Hong et al., 2025). Studies of reasoning capacity and sampling entropy examine which trajectories remain accessible during RL (Yue et al., 2025; Cui et al., 2025b). Thyme (Zhang et al., 2026b) uses different temperatures for reasoning and executable code to improve tool rollouts. ContextRL studies sampling reachability jointly with reward reliability: rare correct responses make false positives especially consequential. Feedback targets all-negative groups; the report ablation uses ordinary sampling with the same rollout budget.

5.2 Verification, Feedback, and Training from Recovered Samples

Process rewards, reasoning decomposition, search, and multi-turn learning enrich supervision and sampled experience (Cui et al., 2025a; Duo et al., 2026; Li et al., 2025; Guan et al., 2025; Wang et al., 2025e); evaluator-bias studies examine the reliability of model judgments (Wang et al., 2024a; 2025a). R1-Reward (Zhang et al., 2026a) trains a multimodal preference evaluator through stabilized RL and consistency rewards. ContextRL provides reference context to an existing evaluator to improve supervision for a task-solving policy. Its contribution within feedback-assisted learning is the linked workflow: recover from groups judged all-negative, verify and retain self-contained successes, then pair those successes and the initial negatives with the original query. This preserves reward contrast while training toward the deployment input. Feedback-conditioned samples introduce a distribution shift; the empirical mixed-group procedure is assessed through final-policy performance. Appendix E provides the broader background.

6 Conclusion

ContextRL addresses reward reliability and sampling reachability as two policy information bottlenecks in RLVR. Reference solutions strengthen verification, while mistake reports recover positives for all-negative groups. Evaluation shows an 8.60% reward-accuracy gain and 8.56% training recovery coverage; component ablations and eleven benchmarks connect these improvements to stronger unaided policies. The Keye-VL 2.0 technical report describes incorporating ContextRL into its reinforcement-learning pipeline (Kwai Keye Team, 2026).

Limitations. Context augmentation supplies additional evidence, but its value depends on the reliability of reference solutions and reports and the policy's ability to use them. Our experiments support improved verification and recovery, while robustness to imperfect context and the conditions under which recovered positives translate into stronger unaided policies remain incompletely characterized. We are refining the experimental setup to examine these questions more broadly. Further large-scale ablation studies are underway to quantify the method's contribution.

LLM USAGE DECLARATION

LLMs were used solely to check grammar and improve the clarity of the manuscript. They did not contribute to the conception of the research or the generation of its substantive content. The authors assume full responsibility for the entire paper, including any language revisions assisted by LLMs. We confirm that this use adheres to applicable ethical standards and does not involve plagiarism or other forms of scientific misconduct.

REFERENCES

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35:23716–23736, 2022.

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-VL: A frontier large vision-language model with versatile abilities. *arXiv preprint arXiv:2308.12966*, 2023.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, et al. Qwen3-VL technical report. *arXiv preprint arXiv:2511.21631*, 2025.

Jiaqi Chen, Jianheng Tang, Jinghui Qin, Xiaodan Liang, Lingbo Liu, Eric Xing, and Liang Lin. GeoQA: A geometric question answering benchmark towards multimodal numerical reasoning. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, August 2021.

Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Jiaqi Wang, Yu Qiao, Dahua Lin, et al. Are we on the right way for evaluating large vision-language models? *arXiv preprint arXiv:2403.20330*, 2024a.

Zhuo Chen, Yichi Zhang, Yin Fang, Yuxia Geng, Lingbing Guo, Xiang Chen, Qian Li, Wen Zhang, Jiaoyan Chen, Yushan Zhu, et al. Knowledge graphs meet multi-modal learning: A comprehensive survey. *arXiv preprint arXiv:2402.05391*, 2024b.

Xianfu Cheng, Wei Zhang, Shiwei Zhang, Jian Yang, Xiangyuan Guan, Xianjie Wu, Xiang Li, Ge Zhang, Jiaheng Liu, Yuying Mai, et al. Simplevqa: Multimodal factuality evaluation for multimodal large language models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4637–4646, 2025.

Krzysztof J Cios, Witold Pedrycz, and Roman W Swiniarski. Data mining and knowledge discovery. In *Data mining methods for knowledge discovery*, pp. 1–26. Springer, 1998.

Ganqu Cui, Lifan Yuan, Zefan Wang, Hanbin Wang, Wendi Li, Bingxiang He, Yuchen Fan, Tianyu Yu, Qixin Xu, Weize Chen, et al. Process reinforcement through implicit rewards. *arXiv preprint arXiv:2502.01456v1*, 2025a. URL https://arxiv.org/abs/2502.01456v1.

Ganqu Cui, Yuchen Zhang, Jiacheng Chen, Lifan Yuan, Zhi Wang, Yuxin Zuo, Haozhan Li, Yuchen Fan, Huayu Chen, Weize Chen, et al. The entropy mechanism of reinforcement learning for reasoning language models. *arXiv preprint arXiv:2505.22617*, 2025b.

Boyang Deng, Songyou Peng, Kyle Genova, Gordon Wetzstein, Noah Snavely, Leonidas Guibas, and Thomas Funkhouser. Visual chronicles: Using multimodal llms to analyze massive collections of images. In *ICCV*, 2025.

Haodong Duan, Xinyu Fang, Junming Yang, Xiangyu Zhao, Zerun Ma, Yuxuan Qiao, Mo Li, Tianhao Liang, Lin Zhu, Amit Agarwal, et al. VLMEvalKit: An open-source toolkit for evaluating large multi-modality models. *arXiv preprint arXiv:2407.11691*, 2024. URL https://arxiv.org/abs/2407.11691.

Jiangshan Duo, Hanyu Li, Hailin Zhang, Yudong Wang, Sujian Li, and Liang Zhao. Judgerlvr: Judge first, generate second for efficient reasoning. *arXiv preprint arXiv:2601.08468*, 2026.

Sebastian Farquhar, Vikrant Varma, Zachary Kenton, Johannes Gasteiger, Vladimir Mikulik, and Rohin Shah. Challenges with unsupervised LLM knowledge discovery. *arXiv preprint arXiv:2312.10029*, 2023.

William J Frawley, Gregory Piatetsky-Shapiro, and Christopher J Matheus. Knowledge discovery in databases: An overview. *AI magazine*, 13(3):57–57, 1992.

Chaoyou Fu, Haojia Lin, Xiong Wang, Yi-Fan Zhang, Yunhang Shen, Xiaoyu Liu, Yangze Li, Zuwei Long, Heting Gao, Ke Li, et al. Vita-1.5: Towards gpt-4o level real-time vision and speech interaction. *arXiv preprint arXiv:2501.01957v1*, 2025. URL https://arxiv.org/abs/2501.01957v1.

Yuxian Gu, Hao Zhou, Fandong Meng, Jie Zhou, and Minlie Huang. Miniplm: Knowledge distillation for pre-training language models. In *ICLR*, 2025.

Tianrui Guan, Fuxiao Liu, Xiyang Wu, Ruiqi Xian, Zongxia Li, Xiaoyu Liu, Xijun Wang, Lichang Chen, Furong Huang, Yaser Yacoob, et al. Hallusionbench: an advanced diagnostic suite for entangled language hallucination and visual illusion in large vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14375–14385, 2024.

Xinyu Guan, Li Lyna Zhang, Yifei Liu, Ning Shang, Youran Sun, Yi Zhu, Fan Yang, and Mao Yang. rstar-math: Small llms can master math reasoning with self-evolved deep thinking. *arXiv preprint arXiv:2501.04519*, 2025.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, et al. DeepSeek-R1 incentivizes reasoning in LLMs through reinforcement learning. *Nature*, 645(8081):633–638, 2025.

Xu Han, Zhengyan Zhang, Ning Ding, Yuxian Gu, Xiao Liu, Yuqi Huo, Jiezhong Qiu, Yuan Yao, Ao Zhang, Liang Zhang, Wentao Han, Minlie Huang, Qin Jin, Yanyan Lan, Yang Liu, Zhiyuan Liu, Zhiwu Lu, Xipeng Qiu, Ruihua Song, Jie Tang, Ji-Rong Wen, Jinhui Yuan, Wayne Xin Zhao, and Jun Zhu. Pre-trained models: Past, present and future. *AI Open*, 2021.

Wenyi Hong, Wenmeng Yu, Xiaotao Gu, Guo Wang, Guobing Gan, Haomiao Tang, Jiale Cheng, Ji Qi, Junhui Ji, Lihang Pan, et al. Glm-4.1v-thinking: Towards versatile multimodal reasoning with scalable reinforcement learning. *arXiv preprint arXiv:2507.01006v1*, 2025. URL https://arxiv.org/abs/2507.01006v1.

Wenxuan Huang, Bohan Jia, Zijie Zhai, Shaosheng Cao, Zheyu Ye, Fei Zhao, Yao Hu, and Shaohui Lin. Vision-r1: Incentivizing reasoning capability in multimodal large language models. *arXiv preprint arXiv:2503.06749v1*, 2025. URL https://arxiv.org/abs/2503.06749v1.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024.

Aniruddha Kembhavi, Mike Salvato, Eric Kolve, Minjoon Seo, Hannaneh Hajishirzi, and Ali Farhadi. A diagram is worth a dozen images. *arXiv preprint arXiv:1603.07396*, 2016.

Kimi Team, Angang Du, Bohong Yin, Bowei Xing, Bowen Qu, Bowen Wang, Cheng Chen, Chenlin Zhang, Chenzhuang Du, Chu Wei, et al. Kimi-VL technical report. *arXiv preprint arXiv:2504.07491*, 2025.

Kwai Keye Team. Kwai keye-vl 1.5 technical report. *arXiv preprint arXiv:2509.01563*, 2025.

Kwai Keye Team. Kwai Keye-VL-2.0 technical report. *ArXiv preprint*, abs/2606.10651, 2026. URL https://arxiv.org/abs/2606.10651.

Xin Lai, Junyi Li, Wei Li, Tao Liu, Tianjian Li, and Hengshuang Zhao. Mini-o3: Scaling up reasoning patterns and interaction turns for visual search. *arXiv preprint arXiv:2509.07969*, 2025.

Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, et al. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*, 2024a.

Haozhan Li, Yuxin Zuo, Jiale Yu, Yuhao Zhang, Zhaohui Yang, Kaiyan Zhang, Xuekai Zhu, Yuchen Zhang, Tianxing Chen, Ganqu Cui, et al. Simplevla-rl: Scaling vla training via reinforcement learning. In *ICLR*, 2026.

Lei Li, Yuqi Wang, Runxin Xu, Peiyi Wang, Xiachong Feng, Lingpeng Kong, and Qi Liu. Multimodal ArXiv: A dataset for improving scientific comprehension of large vision-language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2024b.

Zongxia Li, Wenhao Yu, Chengsong Huang, Rui Liu, Zhenwen Liang, Fuxiao Liu, Jingxi Che, Dian Yu, Jordan Boyd-Graber, Haitao Mi, et al. Self-rewarding vision-language model via reasoning decomposition. *arXiv preprint arXiv:2508.19652v1*, 2025. URL https://arxiv.org/abs/2508.19652v1.

Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, et al. DeepSeek-V3 technical report. *arXiv preprint arXiv:2412.19437*, 2024.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

Guanxing Lu, Wenkai Guo, Chubin Zhang, Yuheng Zhou, Haonan Jiang, Zifeng Gao, Yansong Tang, and Ziwei Wang. Vla-rl: Towards masterful and general robotic manipulation with scalable reinforcement learning. *arXiv preprint arXiv:2505.18719*, 2025a.

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 (ICLR)*, 2024.

Xingyu Lu, Tianke Zhang, Chang Meng, Xiaobei Wang, Jinpeng Wang, Yi-Fan Zhang, Shisong Tang, Changyi Liu, Haojie Ding, Kaiyu Jiang, Kaiyu Tang, Bin Wen, Hai-Tao Zheng, Fan Yang, Tingting Gao, Di Zhang, and Kun Gai. Vlm as policy: Common-law content moderation framework for short video platform. In *SIGKDD*, 2025b.

Linhao Luo, Yuan-Fang Li, Gholamreza Haffari, and Shirui Pan. Reasoning on graphs: Faithful and interpretable large language model reasoning. In *ICLR*, 2024.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, et al. Self-Refine: Iterative refinement with self-feedback. *arXiv preprint arXiv:2303.17651*, 2023. URL https://arxiv.org/abs/2303.17651.

Fanqing Meng, Lingxiao Du, Zongkai Liu, Zhixiang Zhou, Quanfeng Lu, Daocheng Fu, Tiancheng Han, Botian Shi, Wenhai Wang, Junjun He, Kaipeng Zhang, Ping Luo, Yu Qiao, Qiaosheng Zhang, and Wenqi Shao. Mm-eureka: Exploring the frontiers of multimodal reasoning with rule-based reinforcement learning. *arXiv preprint arXiv:2503.07365*, 2025.

Kevin Murphy. Reinforcement learning: an overview. *arXiv preprint arXiv:2412.05265*, 2024.

OpenAI. Introducing openai o3 and o4-mini, 2025. URL https://openai.com/index/introducing-o3-and-o4-mini/.

Shirui Pan, Linhao Luo, Yufei Wang, Chen Chen, Jiapu Wang, and Xindong Wu. Unifying large language models and knowledge graphs: A roadmap. *IEEE Transactions on Knowledge and Data Engineering*, 36(7):3580–3599, 2024. doi: 10.1109/TKDE.2024.3352100.

Runqi Qiao, Qiuna Tan, Guanting Dong, Minhui Wu, Chong Sun, Xiaoshuai Song, Zhuoma GongQue, Shanglin Lei, Zhe Wei, Miaoxuan Zhang, et al. We-math: Does your large multimodal model achieve human-like mathematical reasoning? *arXiv preprint arXiv:2407.01284*, 2024.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, 2021.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, et al. DeepSeekMath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

Noah Shinn, Federico Cassano, Edward Berman, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. Reflexion: Language agents with verbal reinforcement learning. *arXiv preprint arXiv:2303.11366*, 2023. URL https://arxiv.org/abs/2303.11366.

Peiyi Wang, Lei Li, Liang Chen, Zefan Cai, Dawei Zhu, Binghuai Lin, Yunbo Cao, Lingpeng Kong, Qi Liu, Tianyu Liu, and Zhifang Sui. Large language models are not fair evaluators. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*, 2024a.

Qian Wang, Zhanzhi Lou, Zhenheng Tang, Nuo Chen, Xuandong Zhao, Wenxuan Zhang, Dawn Song, and Bingsheng He. Assessing judging bias in large reasoning models: An empirical study. In *COLM*, 2025a.

Wenbin Wang, Liang Ding, Minyan Zeng, Xiabin Zhou, Li Shen, Yong Luo, Wei Yu, and Dacheng Tao. Divide, conquer and combine: A training-free framework for high-resolution image perception in multimodal large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 7907–7915, 2025b.

Xiyao Wang, Zhengyuan Yang, Chao Feng, Hongjin Lu, Linjie Li, Chung-Ching Lin, Kevin Lin, Furong Huang, and Lijuan Wang. Sota with less: Mcts-guided sample selection for data-efficient visual reasoning self-improvement. *arXiv preprint arXiv:2504.07934*, 2025c.

Ye Wang, Qianglong Chen, Zejun Li, Siyuan Wang, Shijie Guo, Zhirui Zhang, and Zhongyu Wei. Simple o3: Towards interleaved vision-language reasoning. *arXiv preprint arXiv:2508.12109*, 2025d.

Zihan Wang, Kangrui Wang, Qineng Wang, Pingyue Zhang, Linjie Li, Zhengyuan Yang, Xing Jin, Kefan Yu, Minh Nhat Nguyen, Licheng Liu, et al. Ragen: Understanding self-evolution in LLM agents via multi-turn reinforcement learning. *arXiv preprint arXiv:2504.20073*, 2025e.

Zirui Wang, Mengzhou Xia, Luxi He, Howard Chen, Yitao Liu, Richard Zhu, Kaiqu Liang, Xindi Wu, Haotian Liu, Sadhika Malladi, Alexis Chevalier, Sanjeev Arora, and Danqi Chen. Charxiv: Charting gaps in realistic chart understanding in multimodal llms. *arXiv preprint arXiv:2406.18521*, 2024b.

Luis Wiedmann, Orr Zohar, Amir Mahla, Xiaohan Wang, Rui Li, Thibaud Frere, Leandro von Werra, Aritra Roy Gosthipaty, and Andrés Marafioti. Finevision: Open data is all you need, 2025. URL https://arxiv.org/abs/2510.17269.

Junkang Wu, Kexin Huang, Jiancan Wu, An Zhang, Xiang Wang, and Xiangnan He. Quantile advantage estimation for entropy-safe reasoning. *arXiv preprint arXiv:2509.22611v1*, 2025. URL https://arxiv.org/abs/2509.22611v1.

Yijia Xiao, Edward Sun, Tianyu Liu, and Wei Wang. Logicvista: Multimodal llm logical reasoning benchmark in visual contexts, 2024. URL https://arxiv.org/abs/2407.04973.

Rui Xue, Xipeng Shen, Ruozhou Yu, and Xiaorui Liu. Efficient end-to-end language model fine-tuning on graphs. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.2*, 2025.

Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Weinan Dai, Tiantian Fan, Gaohong Liu, Lingjun Liu, et al. Dapo: An open-source LLM reinforcement learning system at scale. In *NeurIPS*, 2025.

Yang Yue, Zhiqi Chen, Rui Lu, Andrew Zhao, Zhaokai Wang, Shiji Song, and Gao Huang. Does reinforcement learning really incentivize reasoning capacity in LLMs beyond the base model? In *NeurIPS*, 2025.

Eric Zelikman, Yuhuai Wu, Jesse Mu, and Noah D. Goodman. STaR: Bootstrapping reasoning with reasoning. *arXiv preprint arXiv:2203.14465*, 2022. URL https://arxiv.org/abs/2203.14465.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 11975–11986, 2023.

Kaiyan Zhang, Yuxin Zuo, Bingxiang He, Youbang Sun, Runze Liu, Che Jiang, Yuchen Fan, Kai Tian, Guoli Jia, Pengfei Li, et al. A survey of reinforcement learning for large reasoning models. *arXiv preprint arXiv:2509.08827*, 2025a.

Renrui Zhang, Dongzhi Jiang, Yichi Zhang, Haokun Lin, Ziyu Guo, Pengshuo Qiu, Aojun Zhou, Pan Lu, Kai-Wei Chang, Peng Gao, et al. MathVerse: Does your multi-modal LLM truly see the diagrams in visual math problems? *arXiv preprint arXiv:2403.14624*, 2024. URL https://arxiv.org/abs/2403.14624.

Yi-Fan Zhang, Huanyu Zhang, Haochen Tian, Chaoyou Fu, Shuangqing Zhang, Junfei Wu, Feng Li, Kun Wang, Qingsong Wen, Zhang Zhang, et al. MME-RealWorld: Could your multimodal LLM challenge high-resolution real-world scenarios that are difficult for humans? In *International Conference on Learning Representations*, 2025b.

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, Zhang Zhang, Tingting Gao, Di ZHANG, Guorui Zhou, and Liang Wang. R1-reward: Training multimodal reward model through stable reinforcement learning. In *The Fourteenth International Conference on Learning Representations*, 2026a. URL https://openreview.net/forum?id=4Ewgw9M2xE.

YiFan Zhang, Xingyu Lu, Shukang Yin, Chaoyou Fu, Wei Chen, Xiao Hu, Bin Wen, Kaiyu Jiang, Changyi Liu, Tianke Zhang, Haonan fan, Kaibing Chen, Jiankang Chen, Haojie Ding, Kaiyu Tang, Zhang Zhang, Liang Wang, Fan Yang, Tingting Gao, and Guorui Zhou. Thyme: Think beyond images. In *The Fourteenth International Conference on Learning Representations*, 2026b. URL https://openreview.net/forum?id=gCWLkqK45O.

Shitian Zhao, Haoquan Zhang, Shaoheng Lin, Ming Li, Qilong Wu, Kaipeng Zhang, and Chen Wei. Pyvision: Agentic vision with dynamic tooling. *arXiv preprint arXiv:2507.07998*, 2025a.

Yuze Zhao, Jintao Huang, Jinghan Hu, Xingjun Wang, Yunlin Mao, Daoze Zhang, Zeyinzi Jiang, Zhikai Wu, Baole Ai, Ang Wang, et al. Swift: a scalable lightweight infrastructure for fine-tuning. In *AAAI*, volume 39, pp. 29733–29735, 2025b.

Chujie Zheng, Shixuan Liu, Mingze Li, Xiong-Hui Chen, Bowen Yu, Chang Gao, Kai Dang, Yuqiong Liu, Rui Men, An Yang, et al. Group sequence policy optimization. *arXiv preprint arXiv:2507.18071*, 2025a.

Ziwei Zheng, Michael Yang, Jack Hong, Chenxiao Zhao, Guohai Xu, Le Yang, Chao Shen, and Xing Yu. Deepeyes: Incentivizing" thinking with images" via reinforcement learning. *arXiv preprint arXiv:2505.14362*, 2025b.

Chengke Zou, Xingang Guo, Rui Yang, Junyu Zhang, Bin Hu, and Huan Zhang. DynaMath: A dynamic visual benchmark for evaluating mathematical reasoning robustness of vision language models. In *International Conference on Learning Representations (ICLR)*, 2025.

APPENDIX CONTENTS

A Additional Method Analysis . . . 15
A.1 Training Workflow . . . 15
B Experimental Details . . . 17
C Supplementary Results and Failure Cases . . . 18
D Reward Model Instructions . . . 20
E Extended Related Work . . . 21

A ADDITIONAL METHOD ANALYSIS

A.1 TRAINING WORKFLOW

Algorithm 1 summarizes training. Constant-reward groups provide no reward contrast; mixed groups need a retained positive. Regularization and numerical handling still apply.

Algorithm 1 ContextRL: from augmented context to policy training

Require: Current policy π_θ, verifier (r_ψ, h_ψ), data (x, s), group size G
for each training query (x, s) **do**
  Sample G responses $y_g^{(1)} \sim \pi_\theta(\cdot \mid x)$
  Obtain rewards $R_g^{(1)}$ and reports M_g for failed attempts
  if some $R_g^{(1)} = 1$ **then**
    Use the original-input group $\mathcal{G}_{\text{on}}(x)$ for online training
  else
    Sample $y_g^{(2)} \sim \pi_\theta(\cdot \mid x, y_g^{(1)}, M_g)$
    Retain verifier-accepted, self-contained responses in $\mathcal{P}(x)$
    if $\mathcal{P}(x) \neq \varnothing$ **then**
      Pair retained responses with x and form $\mathcal{G}_{\text{mix}}(x)$
      Scale all mixed-group advantages by λ; mask KL on recovered positives
    end if
  end if
end for

A.2 RLVR INFORMATION FLOW

A dataset provides multimodal queries and reference annotations (x, t). The policy π_θ generates candidate responses, and a verifier $V(x, y, t)$ assigns rewards. In group-relative training, rewards determine relative weights within a sampled group. Conceptually, an advantage-weighted policy update contains terms of the form

$$\widehat{g} = \frac{1}{G} \sum_{k=1}^{G} A_k \nabla_\theta \log \pi_\theta(y_k \mid x). \tag{12}$$

This expression illustrates how sampled responses and verifier-derived advantages enter learning. ContextRL uses one online update per rollout batch, with mixed-group advantage scaling and selective KL regularization as specified in Section 3.3. Recovery samples are generated with feedback and trained under the original query after rollback.

A.3 QUANTIFYING TRAINING-INFORMATION GAINS

We use training-information gains to describe improvements in the reliability and availability of supervision supplied to the policy. A human-annotated evaluation measures the change in reward accuracy; training logs measure the share of queries receiving recovered positives. These assess different operational benefits, with the denominators and units specified below.

A.3.1 Response-Level Reward Diagnostics

For a fixed population of responses with independently verified correctness C, the false-positive and false-negative rates of reward rule j are

$$\alpha_j = \Pr(R_j = 1 \mid C = 0), \qquad \beta_j = \Pr(R_j = 0 \mid C = 1). \tag{13}$$

A difference $\alpha_0 - \alpha_1$ measures a change conditional on an incorrect response. In the 4,000-response human-annotated evaluation (Appendix C.2), final-answer-based and process-aware reward rules correctly classify 3,540 and 3,884 responses, respectively. Their accuracy difference is $(3{,}884 - 3{,}540)/4{,}000 = 0.086$, or **8.60%**. This metric counts correct reward judgments over the full evaluation set, including both correct and incorrect responses.

On this set, the false-positive rate falls from 18.71% to 3.44%, while the false-negative rate falls from 2.62% to 2.23%. Positive-reward precision rises from 80.87% to 95.84%. Equations (2) and (3), derived by total probability and Bayes' rule, distinguish positive reward frequency from positive-label quality.

A.3.2 Query-Level Recovery Diagnostics

Let E_x denote the event that all first-stage rewards are zero, and B_x the event that stage two produces at least one retained response. The fraction of all training queries receiving recovered positives is

$$\rho_{\text{recovered}} = \Pr(E_x \cap B_x) = \Pr(E_x)\Pr(B_x \mid E_x). \tag{14}$$

The reported 8.56% corresponds to the overall-query quantity. With exactly G first-stage responses per query, the recovered groups contain $GQ_{\text{recovered}}$ original responses out of GQ. Their fraction is therefore also 8.56%. Each retained positive makes its initially all-zero reward group mixed, giving the original negatives nonzero reward-relative advantages. This establishes the counting interpretation in Equation (11); it requires a nonempty retained positive set and the described mixed-group normalization. It concerns the reward component before and after recovery, separately from KL regularization or the true correctness of retained samples.

The conditional recovery rate is $\Pr(B_x \mid E_x)$ and additionally requires the number of queries entering stage two. If first-stage group sizes vary, response coverage must be weighted by actual group sizes rather than equated to query coverage.

Interpreting the two measurements. The 8.60% reward-accuracy gain compares two reward rules on a human-annotated response set. The 8.56% recovery coverage counts training queries receiving retained positives. They differ in population, denominator, and meaning, so they are reported separately and are not added into a combined gain. Fixed group size relates recovery coverage to the corresponding first-stage response coverage; it does not equate that coverage to reward-accuracy improvement.

A.3.3 Group Contrast and Information Scope

Under independent binary rewards with fixed per-query positive rate u_x, the probability that a size-G group contains both reward values is

$$\Pr(\text{mixed rewards} \mid x) = 1 - u_x^G - (1 - u_x)^G. \tag{15}$$

This excludes both all-positive and all-negative groups. Reward contrast is necessary for a nonzero group-relative reward advantage but does not guarantee useful supervision: a noisy verifier can create contrast among responses of the same true quality. In a heterogeneous dataset, the expression should be evaluated per query before averaging; substituting an overall mean u need not give the correct group probability.

For nested contexts $T_0 = (x, y, a)$ and $T_1 = (x, y, s)$ with a determined by s, the ideal entropy difference satisfies

$$H(C \mid T_0) - H(C \mid T_1) = I(C; s \mid T_0) \geq 0. \tag{16}$$

Strict reduction requires positive conditional mutual information. Adding written intermediate steps does not automatically ensure this condition, nor does the relation guarantee a finite judge's accuracy. Our reported response and query percentages do not estimate this entropy difference. The practical analysis concerns what context makes accessible to the implemented sampler and verifier.

B Experimental Details

B.1 Dataset Construction

To construct a training dataset with Reference Solution, we select FineVision (Wiedmann et al., 2025) as the data source. The Policy Model used for optimization is Qwen3-VL 8B Instruct, while the Reward Model is Qwen3-VL 32B Instruct. To evaluate ContextRL on both perception and reasoning, we choose VQA and multimodal math problems to conduct experiments. We select several VQA dataset (e.g., ArXivQA (Li et al., 2024b), ThinkLite (Wang et al., 2025c), AI2D (Kembhavi et al., 2016)) and several Math dataset (e.g., MMK12 (Meng et al., 2025), GEOQA (Chen et al., 2021)) from FineVision. The candidate dataset consists of approximately 250,000 data instances each with query and final answer. We then filter all instances based on accuracy (Cui et al., 2025a): For each instance, four responses are generated by the Policy Model, and the Reward Model evaluates their correctness according to the ground truth. Instances with all correct responses are filtered out. The remaining instances are then re-sampled using the stronger Reward Model to generate positive samples. From these positive samples, the Reward Model selects the optimal solution as the reference solution for enhancing reward modeling context. To balance the difficulty, the data is categorized based on the number of correct responses generated by the Policy Model, ensuring an equal number of samples in each category. Ultimately, the training dataset consists of 16K VQA instances and approximately 14K multimodal math instances, each containing a multimodal query, a reference solution, and its final answer.

B.2 Benchmarks

To comprehensively evaluate the effect of ContextRL in enhancing the perception and reasoning capabilities of MLLMs, we select the corresponding two types of benchmarks. For perception evaluation, we choose five benchmarks: SimpleVQA (Cheng et al., 2025), MMStar (Chen et al., 2024a), HallusionBench (Guan et al., 2024), HRBench8K (Wang et al., 2025b), and MME-RealWorld-lite (Zhang et al., 2025b). MMStar primarily evaluates the general capabilities of MLLMs across multiple dimensions, HallusionBench and SimpleVQA focus on assessing hallucinations in generated outputs, while HRBench8K and MME-RealWorld-Lite examine the performance of MLLMs in high-resolution scenarios. For reasoning evaluation, we select six benchmarks: MathVerse (Zhang et al., 2024), MathVista (Lu et al., 2024), LogicVista (Xiao et al., 2024), We-Math (Qiao et al., 2024), CharXiv-RQ (Wang et al., 2024b), and DynaMath (Zou et al., 2025). MathVerse, MathVista, and We-Math target visual mathematical reasoning, emphasizing fine-grained understanding and multi-step reasoning, LogicVista focuses on visually grounded logical reasoning, systematically assessing models' abilities in inductive, deductive, spatial, and symbolic reasoning tasks. CharXiv-RQ evaluates scientific chart reasoning, and DynaMath tests MLLMs' consistency under dynamic visual and textual scenarios.

B.3 Baselines

To compare original-input policy performance under different training methods, we adopt the following baselines:

(1) SFT: We directly fine-tune the policy model using the reference solutions as training targets. This process is equivalent to distilling knowledge from the reward model, which acts as a teacher, combined with rejection sampling on the policy model.

(2) GRPO (Shao et al., 2024): As introduced earlier (Section 2.1), GRPO is the most commonly applied RLVR approach, which improves the accuracy of the policy model by computing intra-group relative advantages and applying policy gradient optimization.

(3) DAPO (Yu et al., 2025): DAPO is a well-known variant of GRPO that enhances training stability and learning efficiency through multiple techniques including clip-shifting, dynamic sampling, token-level policy gradient Loss, and overflowing reward shaping.

(4) Qwen3-VL 32B Instruct: Qwen3-VL 32B Instruct is the largest dense model in the Qwen3-VL open-source family and significantly outperforms Qwen3-VL 8B Instruct across a wide range of tasks.

B.4 Implementation Details

For all our training experiments, we adopt the `Qwen3-VL-8B-Instruct` model as the policy model and `Qwen3-VL-32B-Instruct` as the reward model, with their maximum pixel budgets set to 3M and maximum response lengths capped at 16k tokens. During SFT, the model is trained on 29K samples for 5 epochs, with a learning rate of 1e-5 and a global batch size of 128.

For ContextRL and GRPO, the KL divergence loss coefficient β is set to 0.01. The global batch size is configured to 128, the sampling group size to 8, and the learning rate to 1e-6. All reinforcement learning experiments are conducted for a single training epoch. In both SFT and RL phases, we employ the AdamW optimizer (Loshchilov & Hutter, 2017) with a `cosine_with_min_lr` learning rate scheduler and a warm-up ratio of 0.03. All SFT and RL training is performed on 8 NVIDIA H800 GPUs using the ms-swift framework (Zhao et al., 2025b).

ContextRL makes one policy update per rollout batch. For recovered groups, we multiply all group-relative advantages by $\lambda = 0.01$ and mask the KL term on accepted second-stage responses; the initial negative responses retain KL regularization. Ordinary groups use unscaled advantages.

All the evaluation is carried out via the VLMEvalKit framework (Duan et al., 2024), with GPT-4o (Hurst et al., 2024) judging model generated solutions' correctness. For SFT, we select the checkpoint exhibiting the highest average performance across the 5 training epochs; for RL methods, we directly report results from the single-epoch checkpoint.

C Supplementary Results and Failure Cases

C.1 Additional Details for the SFT Sensitivity Study

Section 4.5 and Table 4 report the SFT experiment varying selected false positives in 50K-example training sets.

Training on selected policy responses and training directly on reference solutions differ in sample count, diversity, filtering, and proximity to the policy distribution. These are plausible explanations for differences in performance, but the experiment does not separate their effects. The baseline We-Math value in this supplementary table differs from the main-table value; the SFT comparisons here use the within-table baseline and are not combined with the main-table results.

C.2 Reward Evaluation on a Mixed Response Set

We evaluate the same 4,000 responses, eight for each of 500 queries, under two reward rules: final-answer-based evaluation and process-aware evaluation. Both use Qwen3-VL-32B-Instruct, giving 8,000 reward judgments. Human annotation partitions the responses into 1,793 with an incorrect final answer, 414 with a correct final answer but a flawed process, and 1,793 that are fully correct. A positive ground-truth label requires both a correct answer and a valid process. Thus, the analysis contains 1,793 positive and 2,207 negative ground-truth labels.

Table 3 reports positive judgments by human-labeled class and pooled response-level metrics. For final-answer-based evaluation, the confusion counts are $\mathrm{TP} = 1{,}746$, $\mathrm{FP} = 413$, $\mathrm{TN} = 1{,}794$, and $\mathrm{FN} = 47$. Process-aware evaluation yields $\mathrm{TP} = 1{,}753$, $\mathrm{FP} = 76$, $\mathrm{TN} = 2{,}131$, and $\mathrm{FN} = 40$. Precision is $\mathrm{TP}/(\mathrm{TP} + \mathrm{FP})$, accuracy is $(\mathrm{TP} + \mathrm{TN})/4{,}000$, the false-positive rate (FPR) is $\mathrm{FP}/2{,}207$, and the false-negative rate (FNR) is $\mathrm{FN}/1{,}793$.

Positive judgments on incorrect responses decrease by 81.60%, from 413 to 76. The largest count reduction occurs among responses with correct final answers but flawed processes: their acceptance rate falls from 91.30% (378/414) to 15.46% (64/414). Meanwhile, acceptance of fully correct responses rises from 97.38% to 97.77%. The observed precision gain therefore accompanies preserved correct-response acceptance, supporting the reward-reliability interpretation in Section 4.4.

The 8.60% accuracy improvement is measured on this 4,000-response evaluation set. It is separate from the recorded epoch's 8.56% recovery coverage, which uses all training queries as its denominator (Appendix A.3). The selected-error detection rates in Table 2 likewise describe a different evaluation population.

C.3 Sampling and Training-Signal Statistics

The recorded training epoch contains approximately 230K first-stage responses (29K queries with eight responses each). The context-augmented reward model labels 51.12%, approximately 118K, as negative. The logs record that 8.56% of training queries obtain retained positives in stage two. With a fixed first-stage group size, those queries contain 8.56% of the original responses, which receive reward-relative contrast after recovery. Appendix A.3 specifies the denominators and explains why the two measured gains are reported separately.

These statistics describe reward assignments and restored training coverage within the recorded epoch. Further efficiency comparisons can relate policy performance to the computation spent on verification, recovery, and optimization.

C.3.1 Recovery on Initially All-Negative Queries

We use Qwen3-VL-8B-Instruct on 200 difficult queries, each with zero positive responses among eight initial samples. Both strategies reuse these responses and draw eight more, for 16 rollouts per query. The second round either uses CAS with mistake reports or continues ordinary sampling. CAS yields 341 positive responses covering 133 queries, versus 57 positives covering 39 queries under ordinary sampling (Table 6). Coverage is the fraction of queries with at least one positive response: 66.5% versus 19.5%, a difference of 47%. Thus, CAS recovers positives for 94 additional queries and yields 5.98 times as many positive responses at the same rollout count. This controls sampling count, while report generation and response lengths can change total computation. These conditional rates concern the initially all-negative subset; the recorded 8.56% recovery statistic uses all training queries. This diagnostic measures sampling reachability, not downstream policy performance.

Table 6: Sampling recovery on 200 initially all-negative queries. Both resampling strategies reuse the initial eight responses and add eight, for 16 rollouts per query. Coverage counts queries with at least one positive.

Sampling	Positive responses	Recovered queries	Coverage (%)
Initial sampling (8)	0	0	0.0
Ordinary resampling	57	39	19.5
CAS	**341**	**133**	**66.5**

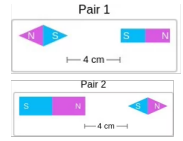
C.4 False-Positive Case Studies

C.4.1 Reasoning Error with a Correct Answer

A correct final answer can coexist with an invalid derivation. Figure A-1 shows a response retrieved from training logs that selects the correct option while misdescribing a magnetic-pole configuration. The highlighted reasoning illustrates an observed false positive for a verifier that checks only the final option. Identifying the inconsistency is relevant to experience discrimination; the example does not establish that training induced reward hacking or that every such error is detected by ContextRL.

C.4.2 Visual Hallucination with a Correct Count

Figure A-2 shows a response retrieved from training logs whose final horse count matches the reference while the reasoning omits a visible horse and describes an unsupported location. This observed example illustrates why a correct total does not validate the observations used to obtain it. It motivates checking the reasoning and visual grounding, without inferring from a static response that the policy deliberately or increasingly exploits the reward.



Context: The images below show two pairs of magnets. The magnets in different pairs do not affect each other. All the magnets shown are made of the same material, but some of them are different sizes and shapes.
Question: Think about the magnetic force between the magnets in each pair. Which of the following statements is true?
Options:
A. The magnitude of the magnetic force is the same in both pairs.
B. The magnitude of the magnetic force is greater in Pair 2.
C. The magnitude of the magnetic force is greater in Pair 1.

Model Response: Let's analyze this step by step.
We are given two pairs of magnets, each separated by 4 cm. We need to compare the **magnitude of the magnetic force** between the magnets in each pair.

Step 1: Understand the setup
- **Pair 1:** The small magnet's N pole faces the bar magnet's S pole → Attraction.
- **Pair 2:** The bar magnet's N pole faces the small magnet's S pole → Attraction.
Both pairs exhibit attractive forces at the same 4 cm distance.

Step 2: Magnetic force dependence
Based on the factors determining magnetic force (strength, distance, orientation):
- **Size & Strength:** Larger magnets generally have stronger magnetic moments.
- **Pair 1:** Involves two smaller magnets.
- **Pair 2:** Involves one large bar magnet and one small magnet.
Since Pair 2 includes a significantly stronger magnet while maintaining the same distance, **the magnetic force in Pair 2 is stronger. **

Step 3: Conclusion
The magnitude of the magnetic force is **greater in Pair 2** because the large magnet in Pair 2 has a stronger magnetic field than either magnet in Pair 1.
✅ Therefore, the correct answer is:
$\boxed{\text{B}}$

Review: The Generated Solution is self-consistent: its reasoning process logically leads to the conclusion that the magnetic force is greater in Pair 2, which matches its final answer. However, it contains a factual error in the description of the pole orientations in Pair 1. The Generated Solution incorrectly states that the N pole of the small magnet faces the S pole of the bar magnet in Pair 1, when in fact, the S pole of the small magnet faces the S pole of the bar magnet (repulsion). Despite this error, the conclusion is correct because the larger magnet in Pair 2 still results in a greater force magnitude. However, since the reasoning contains a factual mistake about the pole alignment, and the Reference Solution correctly identifies the repulsive interaction in Pair 1, the Correctness is rated as 0.

Figure A-1: **False Positive Example (Reasoning Error).**



Question: How many horses are there in the picture?

Model Response:
Identify the animals: Several animals are grazing on a green pasture. Based on their typical horse-like body shape, size, and posture, they appear to be horses. Some are dark-colored; one is light pink (possibly a variation), and some have white markings on the face/legs.
Counting systematically: On the far left of the pasture, there was a black horse with white stripes on its face. To its right, there is another black horse, also grazing. Further to the right, near the fence line, there is a third pink horse grazing. In the lower right corner, partially blocked by the fence, there seems to be a fourth horse, which is also pink.
Check the quantity: Let's count it again to make sure there are no omissions. On the far left: 1 horse. Beside: 1 horse. Beside it is: one pink horse. On the far right: 1 pink horse.
✅ Therefore, the correct answer is:
$\boxed{\text{4}}$

Review: The reasoning process of the generated solution is logically consistent with its final answer, "Four horses". It systematically identified and counted each horse, including the one on the far left and the pink one on the right, which was consistent with the counting result of the reference solution. However, the generated solution wrongly claimed that there was a "fourth horse..." In the lower right corner, part of it is blocked by a fence, which is also pink. This is hallucination - there is no fourth pink horse in the picture at all. The fourth horse is actually the lighter-colored pony on the far left, not the second pink one. Although the final count was correct, this factual error led to flaws in the reasoning.

Figure A-2: **False Positive Example (Hallucination).**

D Reward Model Instructions

Figures A-3 and A-4 give the answer-reference and full-solution reward instructions. Both reject inconsistency between reasoning and the final answer. The context-augmented instruction additionally rejects revision cues so that retained responses remain self-contained after rollback.

Regular Reward Instruction

You are an expert in visual question answering. Given a Question, a Reference Answer, and a Generated Solution (with `reasoning` and a `final answer` in `\boxed{}`), evaluate: (1) Self-Consistency; (2) Correctness vs. the image and Reference Answer; if Correctness=0, write a Mistake Report.
1. **Self-Consistency**:
If the reasoning aligns with the final answer, score 1; otherwise score 0.
2. **Correctness** (only if Self-Consistency=1):
If no hallucinations/conflicts with the image, reasoning is logically correct, and the final answer matches the Reference Answer in meaning, score 1; else 0.
3. **Mistake Report** (only if Self-Consistency=1 and Correctness=0):
Summarize wrong content concisely; do not provide correct facts or reveal the correct final answer.

Inputs:
Question: `<Input Question>`
Reference Answer: `<Ref Answer>`
Generated Solution: `<Gen Solution>`
Output Format (JSON):

Regular Reward Instruction (continued)

```
{
  "Analysis": brief analysis of self-consistency and correctness,
  "Self-Consistency": score1,
  "Correctness": score2,
  "Mistake Report": "mistakes (if applicable)"
}
```

Your Evaluation Result:

Figure A-3: Regular reward instruction template.

Context-Augmented Reward Instruction

You are an expert in visual question answering. Given a Question, a Reference Solution, and a Generated Solution (with `reasoning` and a `final answer` in `\boxed{}`), evaluate Self-Consistency and then Correctness. If Correctness=0, provide a Mistake Report.
1. **Self-Consistency:**
If pleasantries or revision cues (e.g., "You are absolutely right", "Thank you for your review", "I apologize for my mistake", "re-analyze", "reconstruct", "re-examine") appear, set Self-Consistency=0 and Correctness=0; else Self-Consistency=1 iff reasoning matches final answer (otherwise set both to 0).
2. **Correctness** (only if Self-Consistency=1):
Using the image as truth and the Reference Solution as auxiliary, set Correctness=1 iff no hallucinations/conflicts, reasoning is logical, and the final answer matches the Reference Solution's final answer in meaning; else 0.
3. **Mistake Report** (only if Self-Consistency=1 and Correctness=0):
List wrong reasoning content only; **no** evidence, correct facts, or final-answer leakage.

Inputs:
Question: `<Input Question>`
Reference Solution: `<Ref Solution>`
Generated Solution: `<Gen Solution>`
Output Format (JSON):

```
{
  "Analysis": brief analysis of self-consistency and correctness,
  "Self-Consistency": score1,
  "Correctness": score2,
  "Mistake Report": "mistakes (if applicable)"
}
```

Your Evaluation Result:

Figure A-4: Context-augmented reward instruction template.

E Extended Related Work

E.1 Learning from Data and Structured Supervision

Data mining extracts useful patterns and structures from data (Frawley et al., 1992; Cios et al., 1998). Pretrained models encode learned regularities in parameters (Han et al., 2021), while reinforcement learning changes behavior through evaluative interaction (Murphy, 2024).

Complementary directions study language models and knowledge graphs (Pan et al., 2024), faithful graph-based reasoning (Luo et al., 2024), and unsupervised knowledge discovery (Farquhar et al., 2023). Multimodal knowledge-graph learning connects textual and visual evidence (Chen et al., 2024b); graph-aware fine-tuning uses relational structure (Xue et al., 2025). Pretraining distillation transfers supervision between models (Gu et al., 2025), while Visual Chronicles uses multimodal models to analyze large image collections (Deng et al., 2025).

E.2 Multimodal Representation and Post-Training

Feedback also supports iterative improvement. STaR (Zelikman et al., 2022) uses answer-conditioned retries and trains on successful rationales, while Self-Refine (Madaan et al., 2023) and Reflexion (Shinn et al., 2023) use textual feedback to improve later attempts without updating model weights. ContextRL applies feedback during RL training, targeting all-negative groups and pairing verified recoveries with original inputs and initial negatives.

Contrastive image–text learning provides transferable visual representations (Radford et al., 2021; Zhai et al., 2023). Generative vision–language models support few-shot tasks (Alayrac et al., 2022), visual instruction following (Liu et al., 2023), and broader visual understanding (Bai et al., 2023; Li et al., 2024a). Recent pipelines include Qwen3-VL (Bai et al., 2025), Kimi-VL (Kimi Team et al., 2025), Keye-VL 1.5 (Kwai Keye Team, 2025), and GLM-4.1V-Thinking (Hong et al., 2025); VITA-1.5 studies real-time vision and speech interaction (Fu et al., 2025).

Large-scale language-model training (Liu et al., 2024), reasoning models (OpenAI, 2025), and RL surveys (Zhang et al., 2025a) provide post-training foundations. Other applications include interleaved visual reasoning (Wang et al., 2025d), content moderation (Lu et al., 2025b), and robotic manipulation with vision–language–action policies (Li et al., 2026; Lu et al., 2025a).