CAPTION AS REWARD: ENHANCING VISION-LANGUAGE REASONING THROUGH DENSE VISUAL DESCRIPTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Vision-Language Models (VLMs) face challenges in complex visual reasoning tasks, where the contribution of intermediate visual understanding remains underexplored. We present Caption as Reward (CaR), a reinforcement learning framework that evaluates visual understanding quality through its impact on task performance. Unlike approaches that assess visual description quality independently through linguistic metrics, CaR introduces a gain-based reward mechanism that measures how visual descriptions improve task performance relative to direct reasoning. This approach encourages models to adapt their visual understanding strategy to task complexity. We evaluate CaR on eight reasoning benchmarks using Qwen2.5-VL models (3B and 7B parameters). CaR achieves consistent improvements across model scales: our 3B model with 30K training samples reaches 34.2% average accuracy, significantly outperforming both the SFT baseline (22.9% with 20K samples) and the 3B-Instruct baseline (29.8%). Notably, CaR shows substantial improvements over standard supervised fine-tuning, with gains of +11.3 percentage points (34.2% vs 22.9%) on 30K data. For the 7B model, CaR improves performance from 36.5% (GRPO) to 38.1%, a 1.6 percentage point gain, demonstrating robust improvements regardless of model size. CaR's gain-based reward mechanism provides a principled training signal that directly links visual description quality to task performance, opening new directions for improving visual reasoning capabilities in VLMs without requiring expensive human annotations. Additional evaluation on MME-RealWorld confirms CaR's effectiveness in enhancing visual perception abilities, with particularly strong improvements in diagram understanding (+31.4 points) and OCR tasks (+8.1 points).

Keywords: Vision-Language Models, Reinforcement Learning, Visual Reasoning, Caption Generation, Reward Modeling, Multimodal Learning

1 INTRODUCTION

Post-training techniques such as deliberate chain-of-thought reasoning have substantially improved large language models (LLMs) (7; 13). Vision-language models (VLMs) benefit from these ideas, yet accurate perception remains a bottleneck: mainstream systems miss crucial scene details in more than 10% of domain-specific queries (1), and our audit of 1,200 math and science questions attributes 62.1% of failures to incomplete or incorrect visual descriptions.

A common workaround translates images into textual descriptions that are then processed by powerful text-only reasoners (2; 3; 8). Linguistic metrics (e.g., BLEU, ROUGE) may rate such captions highly even when downstream answers do not improve. Reinforcement learning could align captions with task success, but training a bespoke reward model is expensive and brittle (10; 5), whereas rule-based rewards such as those used by DeepSeek-R1 remain surprisingly effective (6).

We introduce *Caption as Reward* (CaR), a lightweight reinforcement learning framework that scores each visual description by the performance gain it unlocks for the base VLM. CaR favors captions that repair direct reasoning failures, penalizes those that degrade correct predictions, and relies only

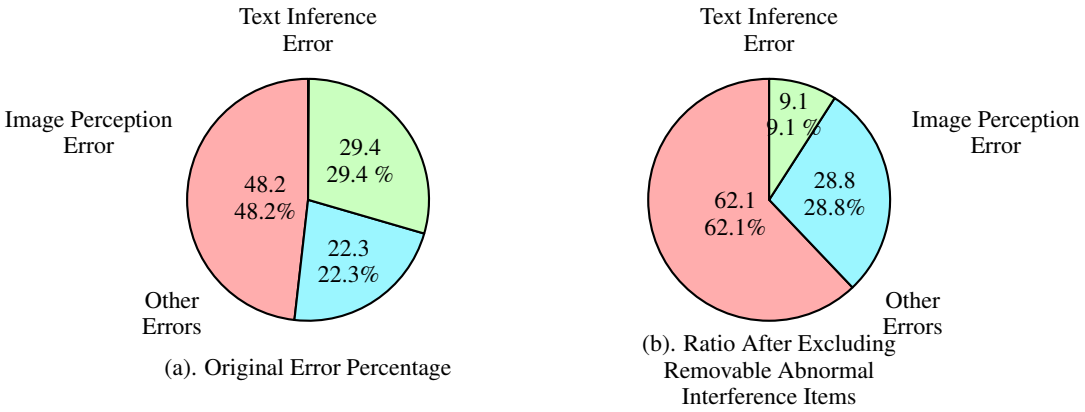

Figure 1: Analysis of the causes of errors in multimodal questions. (a) Original error percentage. (b) Ratio after excluding removable abnormal interference items.

on the existing model plus an external semantic judge. The approach yields consistent gains on eight visual reasoning benchmarks with Qwen2.5-VL backbones of 3B and 7B parameters.

Our contributions are threefold: (i) a gain-based reward formulation that directly measures the utility of a caption for downstream accuracy; (ii) a four-case policy update that plugs into standard GRPO training without auxiliary reward models; and (iii) state-of-the-art accuracy for 3B-scale VLMs together with strong improvements at 7B scale, achieved without additional human annotation.

2 METHODOLOGY

We propose Caption as Reward (CaR), a reinforcement learning framework designed to improve visual reasoning capabilities of VLMs using only visual question-answer pairs without any explicit CoT supervision. In what follows, we first highlight the motivation behind our approach, then introduce our CaR framework which evaluates visual description quality through performance gain rather than linguistic metrics, and finally detail our training objective and implementation.

2.1 MOTIVATION: BEYOND LINGUISTIC VISUAL DESCRIPTION EVALUATION

To understand the dominant failure modes, we manually audited 1,200 multimodal math and science problems with human verification. We examined the subset where Qwen2.5-VL-7B solved the text-only version yet failed once Gemini 2.5 descriptions were routed through DeepSeek-R1. Nearly half of the mistakes (48.2%) were caused by missing or distorted perceptual cues in the captions, compared with 22.3% stemming from textual reasoning and 29.4% from other factors. After filtering ambiguous cases (e.g., vague drawings or incorrect gold labels), perception-related errors climbed to 62.1% (Figure 1). These numbers motivate rewarding captions that genuinely improve visual understanding rather than surface fluency.

Traditional approaches to visual reasoning often rely on sophisticated visual description evaluation metrics such as BLEU, ROUGE, or learned similarity measures. However, these metrics primarily assess linguistic quality rather than the visual description's utility for reasoning tasks. This disconnect becomes problematic when training VLMs for visual reasoning, as the model may learn to generate linguistically fluent but reasoning-irrelevant descriptions.

Our key insight is that visual description quality should be measured by its contribution to downstream reasoning performance. A high-quality visual description should provide sufficient visual information to enable correct reasoning, while a poor visual description should fail to support the reasoning process. This performance-centric view motivates our Caption as Reward framework, where we evaluate visual descriptions not by their linguistic properties, but by their functional utility in visual reasoning tasks.

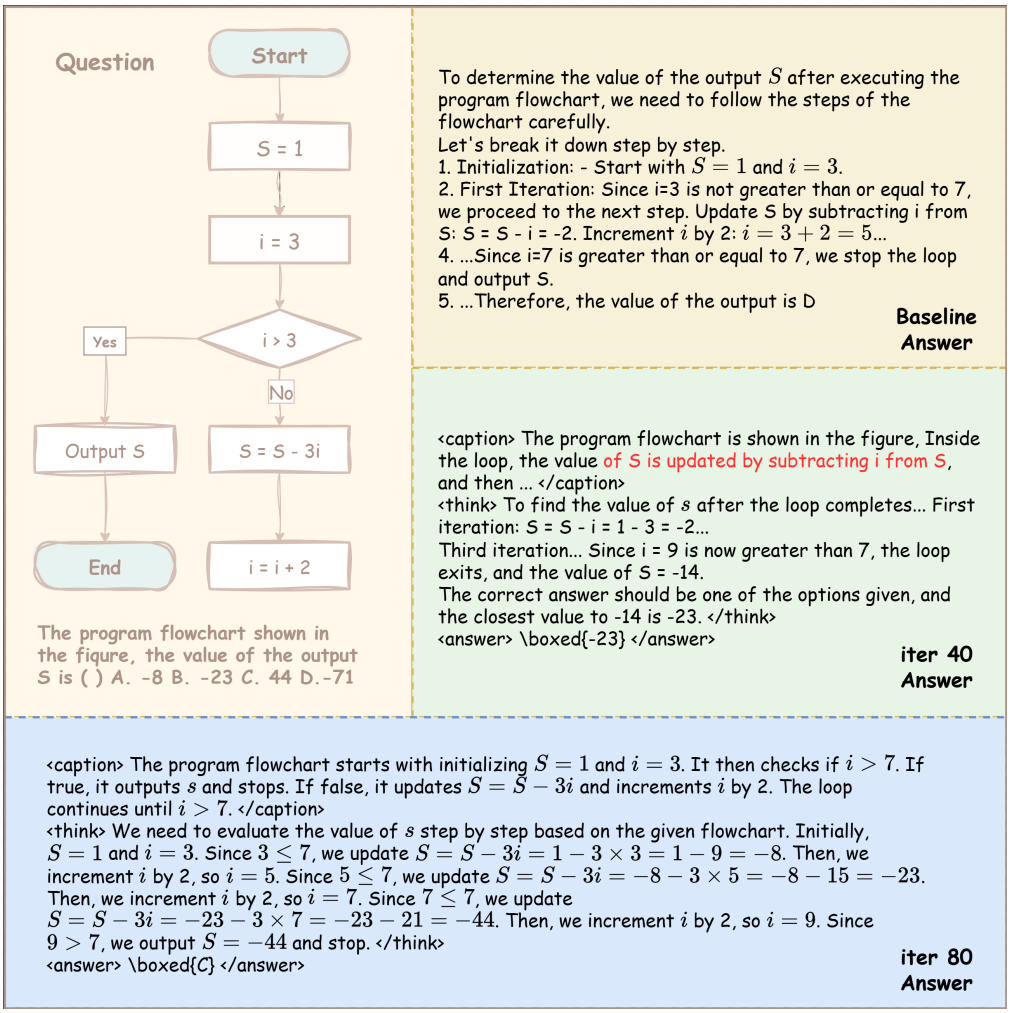

Figure 2: Each column mirrors the VirL39K (left) and MM-Eureka (right) annotation layout: original flowchart on the left, baseline reasoning snapshots in the upper-right blocks, and CaR-enhanced captions in the lower cards. CaR replaces the shortcut baseline with loop-by-loop tracking to recover the correct answer.

2.2 Theoretical Foundation

The effectiveness of CaR can be understood through the lens of information theory. Consider the mutual information between the visual content I and the final answer A:

$$I(I; A) = I(I; C) + I(I; A|C) \tag{1}$$

where C represents the caption. Traditional caption evaluation metrics optimize $I(I; C)$ directly, measuring how much visual information is preserved in the caption. However, CaR optimizes for $I(C; A|Q)$, the information gain that the caption provides for answering the question. This distinction is crucial: a caption may contain abundant visual details (high $I(I; C)$) but fail to include the specific information needed for reasoning (low $I(C; A|Q)$).

Our gain-based reward directly measures this task-relevant information extraction. When the model succeeds with caption but fails without it, we know the caption contains critical visual information that enables correct reasoning. This provides a stronger learning signal than linguistic similarity metrics, which may reward verbose but uninformative descriptions.

2.3 CAPTION AS REWARD DEFINITION

Reward composition. Overall, the reward in this paper consists of three parts: R_{ACC}, R_{format}, and $R(C|I,Q)$, which represent accuracy, format, and visual description gain, respectively. The corresponding weights for these components are W_{acc}, W_{format}, and $W_{caption}$, respectively. The entire reward mechanism can be represented as follows:

$$R = W_{acc} \cdot R_{ACC} + W_{format} \cdot R_{format} + W_{caption} \cdot R(C|I,Q). \tag{2}$$

Here, R_{ACC} is calculated based on the accuracy of the answers generated by the training model. It equals 1 if the answer is correct and 0 if it is incorrect. R_{format} is calculated similarly, with a value of 1 if the response conforms to the target format and 0 otherwise, as shown in the Target Response Format section. The calculation method for $R(C|I,Q)$ is described in the Gain-Based Reward Design section.

Target response format. We follow the structured response template introduced in Appendix **??**, which separates captions, reasoning, and final answers via explicit tags.

Problem formulation. Given a visual reasoning task with input image I, question Q, and ground truth answer A^*, we define two inference paradigms:

- **Direct reasoning:** $A_{direct} \sim P(A|I,Q)$
- **Visual description-enhanced reasoning:** Generate visual description $C \sim P(C|I)$, then $A_{caption} \sim P(A|I,Q,C)$

The core principle is that an effective visual description should improve reasoning performance when direct inference fails, while maintaining accuracy when direct inference succeeds.

Gain-based reward. We evaluate visual description quality through performance gain rather than linguistic metrics. Our reward function is designed as:

$$R(C|I,Q) = \begin{cases} 1.0, & \text{if } A_{caption} = A^* \wedge A_{direct} \neq A^* \\ 0.7, & \text{if } A_{caption} = A^* \wedge A_{direct} = A^* \\ 0.2, & \text{if } A_{caption} \neq A^* \wedge A_{direct} \neq A^* \\ 0, & \text{otherwise} \end{cases} \tag{3}$$

The reward design prioritizes four compact cases: (1) R=1.0 when the caption fixes an error made by direct reasoning; (2) R=0.7 when both paradigms succeed, preserving existing skills without over-rewarding easy instances; (3) R=0.2 when both fail, promoting cautious exploration; and (4) R=0 when the caption harms a previously correct prediction. This signal steers the policy toward descriptions that genuinely elevate task accuracy.

Implementation considerations. Several design choices are critical for CaR's effectiveness:

Inference consistency: We use the same model for both direct and caption-enhanced inference to ensure fair comparison. Using different models could introduce confounding factors.

Temperature control: We set temperature $\tau = 0.9$ during caption generation to encourage diverse visual descriptions while maintaining coherence.

Response format enforcement: Our structured format with explicit tags prevents the model from conflating caption generation with reasoning, ensuring clean separation of visual perception and logical inference.

Full pseudo-code is provided in Appendix 1.

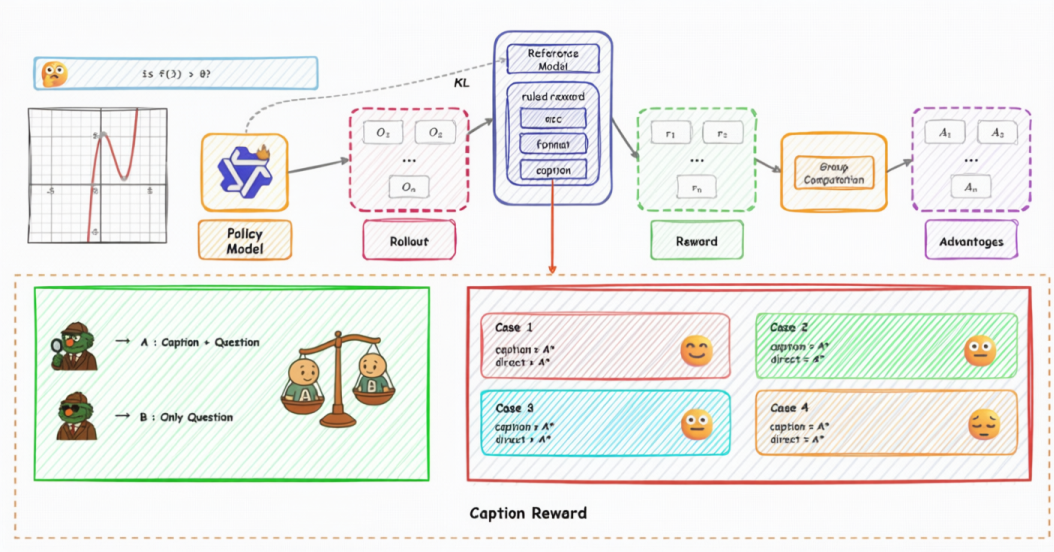

Figure 3: Overall CaR training pipeline. Visual descriptions are scored by their downstream accuracy gain and the policy is updated with GRPO.

2.4 Training Objective and Implementation

We train CaR with group relative policy optimization (GRPO). The training process involves three key stages:

Stage 1: Caption generation and evaluation. For each training sample (I, Q, A^*), we generate $n = 8$ caption candidates using the current policy π_θ. Each caption is evaluated by computing both direct inference A_{direct} and caption-enhanced inference $A_{caption}$, allowing us to measure the caption's contribution to reasoning accuracy.

Stage 2: Reward computation. We compute composite rewards combining three components:

$$R_{total} = w_{acc} \cdot R_{ACC} + w_{format} \cdot R_{format} + w_{caption} \cdot R(C|I, Q) \tag{4}$$

where weights are set to $(1.0, 0.1, 1.0)$ based on ablation studies. The accuracy reward R_{ACC} ensures overall correctness, R_{format} maintains structured output, and $R(C|I, Q)$ provides the caption-specific learning signal.

Stage 3: Policy update. Rewards are normalized within each group of 8 samples:

$$\hat{R}_i = \frac{R_i - \mu_g}{\sigma_g + \epsilon} \tag{5}$$

where μ_g and σ_g are group mean and standard deviation. The policy is then updated using the standard GRPO objective with advantage estimation:

$$\mathcal{L}(\theta) = -\mathbb{E}_{(C,R)\sim\mathcal{D}} \left[\min \left(r_\theta(C)\hat{A}, \text{clip}(r_\theta(C), 1-\varepsilon, 1+\varepsilon)\hat{A} \right) \right] \tag{6}$$

where $r_\theta(C) = \pi_\theta(C|I, Q)/\pi_{\text{ref}}(C|I, Q)$ is the importance ratio and $\varepsilon = 0.2$.

External evaluators (gpt-4o-mini or Qwen2.5-7B-Instruct) provide semantic correctness signals, and a structured response format (Appendix **??**) enforces separate caption/thinking/answer spans. Detailed derivations and computational overhead are provided in Appendix **??**.

3 Related Work

Vision-language model training. Early VLMs relied on supervised fine-tuning with image-text pairs, achieving strong performance on captioning but struggling with complex reasoning tasks.

Recent work explores various post-training strategies to enhance reasoning capabilities. Chain-of-thought distillation (4; 8) transfers reasoning patterns from stronger models but often amplifies perception errors present in the teacher model. Instruction tuning approaches (2) improve task following but lack mechanisms to verify visual understanding quality.

Reinforcement learning for VLMs. Several recent works apply RL to improve VLM performance. Visionary-R1 (12) uses rule-based rewards for self-correction, while VL-Rethinker (11) introduces reflective rewards based on consistency checks. However, these approaches do not explicitly disentangle perception from reasoning improvements. TBAC-VLR1 and VLAA-Thinker focus on action-based rewards but require task-specific reward engineering. CaR's key innovation is using performance gain as a direct measure of caption utility, providing task-agnostic rewards that naturally encourage better visual perception.

Caption evaluation metrics. Traditional caption evaluation relies on n-gram overlap metrics (BLEU, ROUGE, METEOR) or learned similarity measures (BERTScore, CLIPScore). Recent work (10; 9) explores using captions as auxiliary rewards, but still evaluates them through linguistic similarity. CaR fundamentally departs from this paradigm by evaluating captions solely through their contribution to downstream task performance, aligning the training objective directly with the end goal of accurate visual reasoning.

Visual perception in reasoning. Our error analysis revealing 62% perception-related failures aligns with concurrent findings (5) showing that VLMs often fail due to inadequate visual grounding rather than logical errors. Recent benchmarks like MME-RealWorld specifically target perception evaluation, confirming that visual understanding remains a critical bottleneck. CaR directly addresses this challenge by incentivizing captions that capture task-relevant visual information.

4 Experiments

4.1 Experimental Settings

We conduct experiments on a carefully curated training dataset combining MM-Eureka and VirL39K, with comprehensive dataset descriptions, benchmark protocols, and implementation hyperparameters detailed in Appendix **??**. Figure 2 highlights representative prompt–caption pairs from each corpus, illustrating how CaR encourages faithful loop tracing and grounded textual rationales on flowchart-style questions.

4.2 Main Results

Table 1 presents our comprehensive experimental results across 8 diverse benchmarks. Our Caption as Reward method (CaR) demonstrates consistent performance gains over vanilla GRPO across all model configurations. The 3B model with 30K training samples achieves 34.2% average accuracy, significantly outperforming the 3B-Instruct baseline (29.8%) and recent 3B-scale methods including Visionary-R1 (33.0%), TBAC-VLR1 (33.8%), and VLAA-Thinker-3B (32.9%). The 7B model further improves from 36.5% (GRPO) to 38.1% with CaR, demonstrating the scalability of our approach.

4.3 Visual Perception Analysis

To specifically evaluate visual perception capabilities, we conducted additional experiments on MME-RealWorld, a benchmark designed to test fine-grained visual understanding across diverse real-world scenarios. Table 2 shows detailed results across reasoning and perception tasks.

Perception improvements. CaR demonstrates substantial improvements in visual perception tasks, achieving an average score of 53.1 compared to 42.8 for the base model (+10.3 points). Notable gains include diagram understanding (+31.4 points: 71.5 vs 40.1), OCR tasks (+8.1 points: 78.4 vs 70.3), and monitoring tasks (+1.5 points: 33.1 vs 31.6). These results confirm that CaR's gain-based reward mechanism effectively enhances the model's ability to extract relevant visual information for complex reasoning tasks.

Table 1: Comprehensive performance comparison across visual reasoning benchmarks. CaR denotes our Caption as Reward method.

Model	Method	Extra-Model	MMK12	MathVista	MathVision	MathVerse	DynaMath	WeMath	LogicVista	OlympiadBench	Avg
3B Parameter Models with OpenData-10K											
Qwen2.5-VL-3B	SFT	-	29.6	54.7	18.8	8.1	9.7	16.2	34.9	6.0	22.2
Qwen2.5-VL-3B	GRPO	-	48.5	64.1	22.6	32.0	12.5	27.0	41.2	8.0	32.0
Qwen2.5-VL-3B	CaR	gpt-4o-mini	49.8	66.3	22.0	31.6	12.3	27.7	43.4	7.9	32.6
3B Parameter Models with OpenData-20K											
Qwen2.5-VL-3B	SFT	-	33.7	54.0	18.6	9.1	9.2	18.2	34.0	6.0	22.9
3B Parameter Models with OpenData-30K											
Qwen2.5-VL-3B	CaR	gpt-4o-mini	57.6	66.1	25.3	30.1	14.4	28.8	42.1	8.8	34.2
7B Parameter Models with OpenData-20K											
Qwen2.5-VL-7B	GRPO	-	50.8	69.2	25.7	35.9	20.4	36.1	44.3	9.8	36.5
Qwen2.5-VL-7B	CaR	gpt-4o-mini	55.2	70.9	27.2	34.0	20.6	41.1	46.8	9.01	38.1
Baseline Comparisons											
Qwen2.5-VL-3B-Instruct	-	-	41.1	61.2	21.9	31.2	13.2	22.9	40.0	6.8	29.8
Visionary-R1	-	-	45.3	69.4	24.7	33.0	13.8	28.0	41.6	7.8	33.0
TBAC-VLR1	-	-	47.2	64.8	25.0	34.5	17.7	32.4	40.8	8.3	33.8
VLAA-Thinker-3B	-	-	43.2	61.0	24.4	36.4	18.2	33.8	38.5	7.9	32.9

Table 2: Performance on MME-RealWorld benchmark for visual perception evaluation. Results show improvements in both reasoning and perception capabilities.

Model	**Reasoning**					**Perception**					**Avg**
	Monitor	Auto Drive	OCR	Diagram	Remote	Monitor	Auto Drive	OCR	Diagram	Remote	
Qwen2.5-VL-3B-Instruct	22.5	30.0	57.8	46.6	0.0	31.6	35.7	70.3	40.1	22.6	42.8
OpenData-20K (CaR)	21.9	28.2	60.8	52.2	0.0	33.1	37.2	78.4	71.5	26.9	53.1

Qualitative analysis. To understand how CaR improves visual descriptions, we analyzed 100 samples where CaR succeeded but the baseline failed. We identified three key patterns:

1. Enhanced detail extraction: CaR models generate more specific numerical values and spatial relationships. For instance, when counting objects in complex scenes, CaR descriptions explicitly enumerate each visible item rather than providing approximate counts.

2. Task-relevant focus: CaR learns to prioritize information relevant to the question. In mathematical diagrams, CaR descriptions emphasize geometric relationships and measurements while baseline models often describe irrelevant aesthetic features.

3. Systematic coverage: CaR descriptions follow more structured patterns, systematically covering different regions or aspects of the image. This reduces the likelihood of missing critical visual elements.

Error analysis. Despite improvements, CaR still faces challenges in certain scenarios:

Complex spatial reasoning: Tasks requiring 3D understanding or complex spatial transformations remain difficult, as the gain-based reward cannot compensate for fundamental architectural limitations.

Fine-grained recognition: When distinguishing between visually similar objects or symbols, CaR shows limited improvement, suggesting that perception enhancement has diminishing returns for tasks requiring specialized visual expertise.

Compositional reasoning: Multi-step problems requiring both accurate perception and complex reasoning chains show smaller gains, indicating that perception improvements alone cannot solve all reasoning challenges.

4.4 Experimental Details

Datasets. We fine-tune on MM-Eureka and VirL39K. Samples are stratified by difficulty using eight inference runs of Qwen2.5-VL-7B; we drop always-wrong cases and build 10k/20k/30k splits with a 90%/10% mix of medium and high-difficulty items. Figure 2 showcases representative prompts, captions, and reasoning trajectories for the two corpora.

Benchmarks. Evaluation spans eight reasoning benchmarks (MathVista, MathVision, MathVerse, DynaMath, WeMath, LogicVista, MMK12-EVAL, OlympiadBench); we rely on VLMEvalKit except for MMK12-EVAL, where we use the official evaluation script.

Table 3: Impact of different extra-models on visual description reward quality.

Extra-Model	Method	Avg
Qwen2.5-7B-Instruct	CaR	31.6
gpt-4o-mini	CaR	32.6(+1.0)

Models and implementation. All runs use Qwen2.5-VL (3B/7B) with GRPO (n=8, learning rate 5×10^{-7}, temperature 0.9, two epochs); rewards combine accuracy, caption, and format terms with weights $(1.0, 1.0, 0.1)$. Qwen2.5-7B-Instruct or gpt-4o-mini act as evaluators, and training is performed in the verl framework on NVIDIA A100 GPUs.

Computational cost analysis. CaR incurs approximately $3\times$ the computational cost of standard SFT due to:

- Multiple inference passes: 8 caption generations + 16 answer evaluations per sample
- External evaluator calls: Additional API costs for gpt-4o-mini or local inference for Qwen2.5-7B
- GRPO optimization overhead: Advantage estimation and importance sampling computations

However, this cost is justified by significant performance gains (+11.3 points over SFT) and is comparable to other RL-based methods. Future work could explore distillation or reward model caching to reduce computational requirements.

4.5 Ablation Studies

Visual description rewards. We conduct systematic ablation studies to understand the contribution of each component in our reward design. Incorporating visual description rewards consistently improves performance over vanilla GRPO. For the 3B model with 10K data, adding caption rewards increases average accuracy from 32.0% to 32.6%. This gain becomes more pronounced with larger datasets: with 20K samples, the improvement grows from 32.7% to 33.7%, and with 30K samples, we achieve our best result of 34.2%.

The performance gains are particularly notable on benchmarks requiring detailed visual understanding. On MMK12-Eval, the 3B model improves from 48.0% (GRPO) to 57.6% (CaR) with 20K data—a remarkable 20% relative improvement. This suggests that our caption reward mechanism effectively addresses the visual perception bottleneck in complex reasoning tasks.

Data scale. We investigate how our method performs with different amounts of training data. The results demonstrate a clear scaling trend: 10K data yields 32.6% average accuracy, 20K achieves 33.7%, and 30K reaches the best performance at 34.2%. This consistent improvement indicates that our reward mechanism effectively leverages additional training data to enhance visual reasoning capabilities.

Auxiliary evaluators. We investigated the impact of using different evaluation models for visual description reward calculation. We employed two distinct auxiliary models, namely qwen2.5-7b-instruct and gpt-4o-mini, with training data sourced from OpenData-10K. As shown in Table 3, the results indicate that utilizing gpt-4o-mini yielded more significant performance gains, suggesting that a more robust auxiliary model introduces less evaluation noise and more directly reflects the influence of the model's visual descriptions on problem resolution.

5 Discussion and Limitations

Why does CaR work? Our analysis suggests three factors contribute to CaR's effectiveness:

1. Direct optimization for task utility: Unlike methods that optimize proxy metrics (linguistic similarity, rule compliance), CaR directly rewards captions that improve task performance. This creates a tight feedback loop between perception quality and reasoning accuracy.

2. Implicit curriculum learning: The gain-based reward naturally creates a curriculum where the model first learns to fix obvious perception errors (high reward), then gradually improves on subtler cases. This organic difficulty progression may explain why CaR scales better with data than SFT.

3. Disentangled learning signals: By separately evaluating direct and caption-enhanced reasoning, CaR provides clearer gradient signals about what visual information is missing. This helps the model learn which visual features are task-relevant rather than memorizing caption patterns.

Limitations and future directions. While CaR demonstrates strong results, several limitations warrant discussion:

Computational overhead: The $3\times$ training cost compared to SFT may limit adoption for resource-constrained settings. Future work should explore more efficient reward computation strategies, such as caching evaluator responses or using lightweight reward models.

Single architecture evaluation: We evaluated CaR only on Qwen2.5-VL models. Testing on diverse architectures (LLaVA, BLIP, Flamingo) would strengthen claims about generalizability.

Limited to visual QA: Current experiments focus on question-answering tasks. Extending CaR to other modalities (video, audio) and tasks (generation, editing) remains unexplored.

Reward design choices: Our reward weights and thresholds were determined through limited grid search. More principled approaches using multi-objective optimization or learned reward functions could improve performance.

Broader impacts. CaR's improved visual perception could enable more reliable VLM deployments in education, accessibility, and scientific research. However, enhanced visual understanding also raises concerns about potential misuse for surveillance or generating misleading content. We recommend careful deployment with appropriate safeguards and regular auditing of model outputs.

6 Conclusion

We introduced Caption as Reward (CaR), a lightweight reinforcement learning framework that scores visual descriptions by the accuracy gains they unlock. Across eight benchmarks, CaR demonstrates substantial improvements over both supervised fine-tuning and existing reinforcement learning approaches: at 3B scale, CaR achieves 34.2% average accuracy compared to 22.9% for SFT and 29.8% for the base model, representing gains of +11.3 and +4.4 percentage points respectively. At 7B scale, CaR improves from 36.5% (GRPO) to 38.1%, outperforming recent reinforcement-learning baselines without bespoke reward models. The main takeaway is that performance-aligned rewards provide a sharper learning signal than linguistic metrics, enabling smaller VLMs to close part of the gap to larger systems. Current limitations include evaluation on a single backbone family, a $3\times$ training cost relative to supervised fine-tuning, and heuristic reward weights; future work will target broader architectures, efficiency improvements, and multimodal tasks beyond visual question answering.

Ethics Statement

CaR targets educational and scientific reasoning workloads by strengthening factual visual understanding. Although stronger perception could be misused to generate misleading analyses, the method reduces hallucinated descriptions and relies only on public datasets, which we acknowledge may carry existing societal biases. The $3\times$ training cost versus supervised fine-tuning should be weighed against the accuracy gains when deploying the approach. We adhere to the ICLR Code of Ethics and confirm that our work complies with all ethical guidelines. All authors of this work have read and commit to adhering to the ICLR Code of Ethics.

Reproducibility Statement

We train on the publicly available MM-Eureka and VirL39K corpora using Qwen2.5-VL models (3B/7B). All hyperparameters (GRPO with n=8, learning rate 5×10^{-7}, temperature 0.9, two epochs, reward weights $(1.0, 1.0, 0.1)$) and evaluator choices (gpt-4o-mini or Qwen2.5-7B-Instruct) are described in Section 4.1.3. Code, data splits, and checkpoints will be released upon acceptance.

References

[1] Jiaxin Ai, Pengfei Zhou, Zhaopan Xu, Ming Li, Fanrui Zhang, Zizhen Li, Jianwen Sun, Yukang Feng, Baojin Huang, Zhongyuan Wang, and Kaipeng Zhang. Projudge: A multi-modal multi-discipline benchmark and instruction-tuning dataset for mllm-based process judges. *arXiv preprint arXiv:2503.06553*, 2025. URL https://arxiv.org/abs/2503.06553.

[2] Hardy Chen, Haoqin Tu, Fali Wang, Hui Liu, Xianfeng Tang, Xinya Du, Yuyin Zhou, and Cihang Xie. SFT or RL? an early investigation into training r1-like reasoning large vision-language models. *arXiv preprint arXiv:2504.11468*, 2025. URL https://arxiv.org/abs/2504.11468.

[3] Yihe Deng, Hritik Bansal, Fan Yin, Nanyun Peng, Wei Wang, and Kai-Wei Chang. Openvlthinker: Complex vision-language reasoning via iterative sft-rl cycles. *arXiv preprint arXiv:2503.17352*, 2025. URL https://arxiv.org/abs/2503.17352.

[4] Yifan Du, Zikang Liu, Yifan Li, Wayne Xin Zhao, Yuqi Huo, Bingning Wang, Weipeng Chen, Zheng Liu, Zhongyuan Wang, and Ji-Rong Wen. Virgo: A preliminary exploration on reproducing o1-like MLLM. *arXiv preprint arXiv:2501.01904*, 2025. URL https://arxiv.org/abs/2501.01904.

[5] Adam Gleave and Geoffrey Irving. Uncertainty estimation for language reward models. *arXiv preprint arXiv:2203.07472*, 2022. URL https://arxiv.org/abs/2203.07472.

[6] Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Zhi-Feng Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, and Aixin Liu. Deepseek-r1: Incentivizing reasoning capability in LLMs via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025. URL https://arxiv.org/abs/2501.12948.

[7] Komal Kumar, Tajamul Ashraf, Omkar Thawakar, Rao Muhammad Anwer, Hisham Cholakkal, Mubarak Shah, Ming-Hsuan Yang, Phillip H. S. Torr, Fahad Shahbaz Khan, and Salman Khan. Llm post-training: A deep dive into reasoning large language models. *arXiv preprint arXiv:2502.21321*, 2025. URL https://arxiv.org/abs/2502.21321.

[8] Yuan-Hong Liao, Sven Elflein, Liu He, Laura Leal-Taixé, Yejin Choi, Sanja Fidler, and David Acuna. Longperceptualthoughts: Distilling system-2 reasoning for system-1 perception. *arXiv preprint arXiv:2504.15362*, 2025. URL https://arxiv.org/abs/2504.15362.

[9] Yi Peng, Peiyu Wang, Xiaokun Wang, Yichen Wei, Jiangbo Pei, Weijie Qiu, Ai Jian, Yunzhuo Hao, Jiachun Pan, Tianyidan Xie, Li Ge, Rongxian Zhuang, Xuchen Song, Yang Liu, and Yahui Zhou. Skywork r1v: Pioneering multimodal reasoning with chain-of-thought. *arXiv preprint arXiv:2504.05599*, 2025. URL https://arxiv.org/abs/2504.05599.

[10] Juan Rocamonde, Victoriano Montesinos, Elvis Nava, Ethan Perez, and David Lindner. Vision-language models are zero-shot reward models for reinforcement learning. In *Proceedings of the International Conference on Learning Representations*, 2024. URL https://arxiv.org/abs/2310.12921.

[11] Haozhe Wang, Chao Qu, Zuming Huang, Wei Chu, Fangzhen Lin, and Wenhu Chen. Vl-rethinker: Incentivizing self-reflection of vision-language models with reinforcement learning. *arXiv preprint arXiv:2504.08837*, 2025. URL https://arxiv.org/abs/2504.08837.

[12] Yufei Zhan, Yousong Zhu, Shurong Zheng, Hongyin Zhao, Fan Yang, Ming Tang, and Jinqiao Wang. Vision-r1: Evolving human-free alignment in large vision-language models via vision-guided reinforcement learning. *arXiv preprint arXiv:2503.18013*, 2025. URL https://arxiv.org/abs/2503.18013.

[13] Shengjia Zhang, Junjie Wu, Jiawei Chen, Changwang Zhang, Xingyu Lou, Wangchunshu Zhou, Sheng Zhou, Can Wang, and Jun Wang. Othink-r1: Intrinsic fast/slow thinking mode switching for over-reasoning mitigation. *arXiv preprint arXiv:2506.02397*, 2025. URL https://arxiv.org/abs/2506.02397.

A Appendix

B Gain-Based Reward Implementation

Algorithm 1 summarizes the reward computation pipeline referenced in Section 2.

Algorithm 1 Gain-Based Reward Computation

Require: Image I, Question Q, Generated Caption C, Ground Truth A^*, Vision-Language Model $\mathcal{M}$, External Evaluator $\mathcal{E}$
Ensure: Gain-based reward $R(C|I, Q)$
$A_{\text{direct}} \leftarrow \mathcal{M}(I, Q)$ {Direct reasoning}
$A_{\text{caption}} \leftarrow \mathcal{M}(I, Q, C)$ {Caption-enhanced reasoning}
$\text{correct}_{\text{direct}} \leftarrow \mathcal{E}.\text{SemanticMatch}(A_{\text{direct}}, A^*)$
$\text{correct}_{\text{caption}} \leftarrow \mathcal{E}.\text{SemanticMatch}(A_{\text{caption}}, A^*)$
if $\text{correct}_{\text{caption}}$ **and** $\neg\text{correct}_{\text{direct}}$ **then**
  $R(C|I, Q) \leftarrow 1.0$ {Caption fixes an error}
else if $\text{correct}_{\text{caption}}$ **and** $\text{correct}_{\text{direct}}$ **then**
  $R(C|I, Q) \leftarrow 0.7$ {Caption confirms success}
else if $\neg\text{correct}_{\text{caption}}$ **and** $\neg\text{correct}_{\text{direct}}$ **then**
  $R(C|I, Q) \leftarrow 0.2$ {Both attempts fail}
else
  $R(C|I, Q) \leftarrow 0.0$ {Caption harms accuracy}
end if
return $R(C|I, Q)$