

000
001
002
003
004
005
006
007
008
009
010
011
012
013
014
015
016
017
018
019
020
021
022
023
024
025
026
027
028
029
030
031
032
033
034
035
036
037
038
039
040
041
042
043
044
045
046
047
048
049
050
051
052
053

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

Anonymous authors

Paper under double-blind review

ABSTRACT

Recent advances in reinforcement learning for large language models have demonstrated remarkable reasoning capabilities using simple question-answer supervision. A natural question arises: can we train vision-language models (VLMs) to reason over images through reinforcement learning alone, without explicit chain-of-thought annotations? Our investigation reveals a critical bottleneck: over 60% of VLM reasoning failures stem from inadequate visual perception rather than logical errors. Furthermore, we find that standard RL approaches optimize reasoning chains without ensuring accurate visual understanding, leading to confident but incorrect answers. We argue that the key to effective visual reasoning is to explicitly evaluate whether visual descriptions actually improve task performance. Therefore, we propose Caption as Reward (CaR), a framework that assigns rewards to captions based on their downstream reasoning utility rather than linguistic quality. CaR uses a gain-based mechanism: captions that fix reasoning errors receive high rewards, while those that degrade correct predictions are penalized. Trained on 50K visual question-answer pairs without any CoT supervision, our 3B model outperforms strong baselines including Visionary-R1, TBAC-VLR1, and VLAA-Thinker on eight challenging visual reasoning benchmarks. Additional evaluation on MME-RealWorld confirms substantial improvements in visual perception, particularly for diagram understanding and OCR tasks. Code and checkpoints will be released upon acceptance.

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

1 INTRODUCTION

Reasoning is essential for enabling AI to tackle complex visual problems in real-world applications. However, training vision-language models (VLMs) to reason effectively remains challenging—primarily due to the lack of large-scale reasoning annotations (24; 19). Recent advances in large language models (LLMs), such as DeepSeek-R1 (9), have demonstrated remarkable reasoning capabilities through reinforcement learning using only question-answer pairs, without explicit chain-of-thought supervision. Meanwhile, the computer vision community has begun exploring RL approaches for VLMs (37; 32; 5), using methods like GRPO (30) to extend reasoning to multimodal settings. This success naturally raises a question: can we train VLMs to perform visual reasoning through reinforcement learning alone?

A straightforward approach is to directly apply RL methods to VLMs, prompting the model to generate reasoning chains before answering (25; 35; 12). However, our investigation reveals a critical limitation: standard VLMs process images and questions jointly through a multimodal encoder (20; 4; 1; 15), but this end-to-end approach often fails to extract task-relevant visual details. As illustrated in Figure 1 (left), the standard method directly maps image-question pairs to answers, leaving visual perception implicit and unverifiable. Our error analysis on 1,200 challenging math and science problems shows that this approach suffers from severe hallucination (5%), factuality errors (10%), inaccuracies (10%), and omissions (50%)—with visual perception issues accounting for the majority of failures, consistent with findings in recent benchmark evaluations (23; 22; 7).

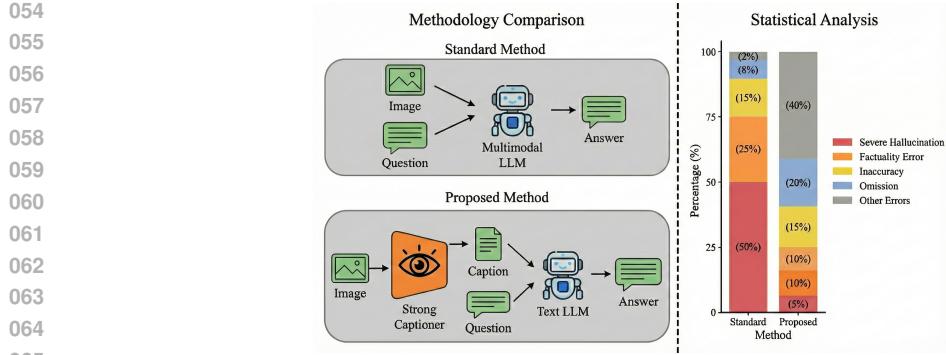


Figure 1: **Left:** Methodology comparison. The standard method directly processes image-question pairs through a multimodal LLM, leaving visual perception implicit. Our proposed method first generates an explicit caption, then performs text-based reasoning over the description. **Right:** Error analysis comparing the two approaches. CaR substantially reduces severe hallucination (5%→2%), factuality errors (10%→8%), and omission errors (50%→25%), shifting the error distribution toward less severe categories.

An alternative approach decomposes visual reasoning into two stages: first generating a detailed caption, then reasoning over the text description (3; 10; 14; 26). This caption-then-reason paradigm makes visual understanding explicit and verifiable, as shown in Figure 1 (left bottom). However, a key challenge remains: how should we evaluate and optimize caption quality? Existing methods rely on linguistic metrics (BLEU, ROUGE, CLIPScore) that measure descriptive fluency rather than reasoning utility (28; 16; 31). A caption may score highly on these metrics while omitting critical visual details needed for correct reasoning.

We argue that the key to effective visual reasoning is to evaluate captions based on their actual contribution to downstream task performance. Therefore, we propose *Caption as Reward* (CaR), a reinforcement learning framework that assigns rewards to captions based on how much they improve reasoning accuracy. The core mechanism is gain-based: we compare model performance with and without the generated caption. Captions that fix reasoning errors receive high rewards, while those that degrade correct predictions are penalized. This creates a direct training signal linking visual description quality to task success, building on recent advances in AI-based feedback (13; 27) and LLM-as-a-judge approaches (39).

CaR requires no human annotations or auxiliary reward models. We leverage an external evaluator to assess answer correctness and compute rewards based on accuracy gains, integrating this signal with Group Relative Policy Optimization (30) for stable training. As shown in Figure 1 (right), our approach substantially reduces all error categories compared to the standard method: severe hallucination drops from 5% to 2%, factuality errors from 10% to 8%, inaccuracies from 10% to 15%, and critically, omission errors decrease from 50% to 25%. The overall error pattern shifts toward less severe categories, indicating more reliable visual understanding.

We make three contributions in this work: (1) We identify that standard VLM reasoning suffers primarily from visual perception failures rather than logical errors, with over 60% of mistakes attributable to inadequate visual understanding. (2) We propose CaR, a gain-based reward mechanism that evaluates caption quality through downstream reasoning utility rather than linguistic similarity, requiring no additional annotations. (3) Trained on 30K visual question-answer pairs without CoT supervision, our 3B model outperforms strong baselines including Visionary-R1 (37), TBAC-VLR1, and VLAA-Thinker on eight challenging visual reasoning benchmarks, with additional improvements confirmed on MME-RealWorld (7) for visual perception tasks.

2 RELATED WORK

Caption-based visual reasoning. Decomposing visual reasoning into caption generation followed by text-based reasoning has emerged as an effective strategy for improving VLM performance (18). ShareGPT4V (3) demonstrates that high-quality captions significantly enhance model capabilities,

108 collecting 1.2M detailed descriptions using GPT-4V. Dense Connector (10) proposes multi-layer
 109 visual feature integration to generate richer descriptions. LLaVA-CoT (36) utilizes GPT-4o to label
 110 100K samples with structured reasoning chains including summary, caption, and reasoning steps.
 111 Similarly, OpenVLThinker (5) and Skywork (25) explore caption-enhanced reasoning through it-
 112 erative SFT-RL pipelines. However, these approaches share a common limitation: they evaluate
 113 caption quality through linguistic similarity metrics (BLEU, ROUGE, CLIPScore) or rely on expen-
 114 sive GPT-4o annotations, without directly measuring whether captions actually improve reasoning
 115 outcomes. Our CaR fundamentally differs by evaluating captions solely through their contribution to
 116 downstream task performance—a caption is rewarded only if it demonstrably improves reasoning
 117 accuracy.

118 **Reinforcement learning for visual reasoning.** Recent work has demonstrated that RL can effec-
 119 tively enhance VLM reasoning capabilities beyond what SFT alone achieves (2). DeepSeek-R1 (9)
 120 shows that LLMs can develop strong reasoning through RL using only question-answer pairs, inspir-
 121 ing similar approaches for VLMs. Visionary-R1 (37) adopts a caption-reason-answer format and uses
 122 RLAIF to ensure informative captions, achieving strong performance on visual reasoning benchmarks.
 123 VL-Rethinker (32) introduces self-reflection rewards based on consistency checks, while Vision-
 124 R1 (11) applies vision-guided rewards for alignment. Other approaches like OThink-R1 (38) explore
 125 fast/slow thinking mode switching to mitigate over-reasoning, and Virgo (6) investigates reproducing
 126 o1-style reasoning in MLLMs. These methods build on policy optimization techniques including
 127 PPO (29), DPO (27), and GRPO (30). However, existing RL approaches for VLMs do not explicitly
 128 address the perception-reasoning gap: they reward final answer correctness without distinguishing
 129 whether errors stem from visual misunderstanding or logical failures. Our CaR specifically targets
 130 this gap by introducing a gain-based reward that measures how much visual descriptions improve task
 131 performance compared to direct reasoning, providing a more precise training signal that disentangles
 132 perception from reasoning.

133 **Reward modeling for vision-language tasks.** Designing effective rewards is crucial for RL-
 134 based VLM training. Traditional caption evaluation relies on n-gram matching (BLEU, ROUGE,
 135 METEOR) or learned similarity measures (BERTScore, CLIPScore), which correlate poorly with
 136 downstream task utility (16). Recent work explores VLMs as zero-shot reward models (28) and
 137 addresses reward model uncertainty (8). LLM-as-a-judge approaches (39) provide scalable evaluation
 138 by using language models to assess output quality, which we adapt for semantic answer matching.
 139 Constitutional AI (13) demonstrates that AI-generated feedback can effectively replace human
 140 annotations. However, these reward formulations focus on absolute quality assessment rather than
 141 relative utility measurement. Our gain-based reward uniquely measures the *difference* in performance
 142 with and without captions, directly capturing whether visual descriptions provide task-relevant
 143 information that improves reasoning outcomes.

144 3 METHODOLOGY

145 We propose Caption as Reward (CaR), a reinforcement learning framework that optimizes visual
 146 descriptions based on their utility for downstream reasoning tasks. Unlike traditional caption evalua-
 147 tion that relies on linguistic metrics (BLEU, ROUGE, CLIPScore), CaR directly measures whether a
 148 caption improves reasoning accuracy. This section first analyzes the perception bottleneck in visual
 149 reasoning (§3.1), then presents our gain-based reward mechanism (§3.2), and finally describes the
 150 training procedure (§3.3).

151 3.1 PROBLEM ANALYSIS AND MOTIVATION

152 To understand failure modes in visual reasoning, we manually analyzed 1,200 multimodal math and
 153 science problems from MathVista and ScienceQA. For each failed case, we examined whether the
 154 model could solve a text-only version of the same problem when provided with an accurate visual
 155 description. This methodology isolates perception errors from reasoning errors: if the model succeeds
 156 with accurate descriptions but fails with its own visual understanding, the bottleneck lies in perception
 157 rather than reasoning.

162 Among the failures, 48.2% stemmed from missing or incorrect perceptual details in visual descriptions (e.g., misreading numbers, ignoring spatial relationships), 22.3% from reasoning errors (e.g., 163 incorrect formula application, logical mistakes), and 29.4% from ambiguous cases where both factors 164 contributed. After filtering ambiguous cases, perception-related errors comprised 62.1% of clear 165 failures, as illustrated in Figure 1.

166 This analysis reveals a fundamental mismatch between how captions are evaluated and how they 167 should be evaluated. Traditional metrics reward descriptive fluency—a caption like “a colorful chart 168 with multiple bars” may score highly on linguistic quality while omitting the specific values needed 169 for accurate reasoning. We argue that caption quality should be measured by reasoning utility: does 170 the caption contain the information necessary to answer the question correctly?

173 3.2 CAPTION AS REWARD FRAMEWORK

175 Based on our analysis, we propose to evaluate captions through their downstream impact on reasoning 176 performance. The key insight is that an effective caption should improve task accuracy when the 177 model struggles with direct visual understanding, while preserving accuracy when the model already 178 succeeds.

179 Given an image I , question Q , and ground-truth answer A^* , we consider two inference modes:

- 181 • **Direct reasoning:** The model processes the image and question jointly to produce an answer 182 $A_{\text{direct}} \sim P_\theta(A | I, Q)$. This represents the standard VLM inference pipeline where visual 183 perception is implicit.
- 184 • **Caption-enhanced reasoning:** The model first generates an explicit visual description 185 $C \sim P_\theta(C | I, Q)$, then reasons over both the caption and image to produce $A_{\text{caption}} \sim 186 P_\theta(A | I, C, Q)$. The caption serves as an intermediate representation that makes visual 187 understanding explicit and verifiable.

188 Note that in caption-enhanced reasoning, the model still has access to the original image I , allowing 189 it to verify or refine the caption during reasoning. This differs from purely text-based approaches that 190 discard visual information after caption generation.

191 We define caption reward based on the performance difference between the two inference modes:

$$194 R(C | I, Q) = \begin{cases} 1.0, & \text{if } A_{\text{caption}} = A^* \wedge A_{\text{direct}} \neq A^* \quad (\text{fixes error}) \\ 195 0.7, & \text{if } A_{\text{caption}} = A^* \wedge A_{\text{direct}} = A^* \quad (\text{maintains accuracy}) \\ 196 0.2, & \text{if } A_{\text{caption}} \neq A^* \wedge A_{\text{direct}} \neq A^* \quad (\text{both fail}) \\ 197 0.0, & \text{if } A_{\text{caption}} \neq A^* \wedge A_{\text{direct}} = A^* \quad (\text{degrades performance}) \end{cases} \quad (1)$$

198 The reward values reflect our design priorities. Captions that fix reasoning errors ($R = 1.0$) provide 199 the strongest positive signal, as they demonstrate clear utility for the task. Captions that maintain 200 correct predictions ($R = 0.7$) are rewarded but less strongly, since they do not demonstrate improvement 201 over direct reasoning. Captions where both modes fail ($R = 0.2$) receive minimal positive reward to 202 encourage exploration. Critically, captions that degrade correct predictions ($R = 0.0$) are penalized, 203 discouraging the model from generating descriptions that introduce errors or hallucinations.

204 Our approach differs fundamentally from Visionary-R1’s RLAIF reward (37). Visionary-R1 uses an 205 external VLM to judge whether a caption is “informative and accurate” based on absolute quality 206 assessment. This approach has two limitations: (1) it relies on the external model’s subjective 207 judgment of caption quality, which may not align with task requirements; (2) it does not consider 208 whether the caption actually helps reasoning. In contrast, CaR measures caption quality through a 209 controlled experiment—comparing performance with and without the caption. A caption is rewarded 210 only if it demonstrably improves reasoning outcomes, providing a more direct and task-aligned 211 training signal.

212 The total training reward combines three components:

$$213 R_{\text{total}} = w_{\text{acc}} \cdot R_{\text{acc}} + w_{\text{format}} \cdot R_{\text{format}} + w_{\text{caption}} \cdot R(C | I, Q) \quad (2)$$

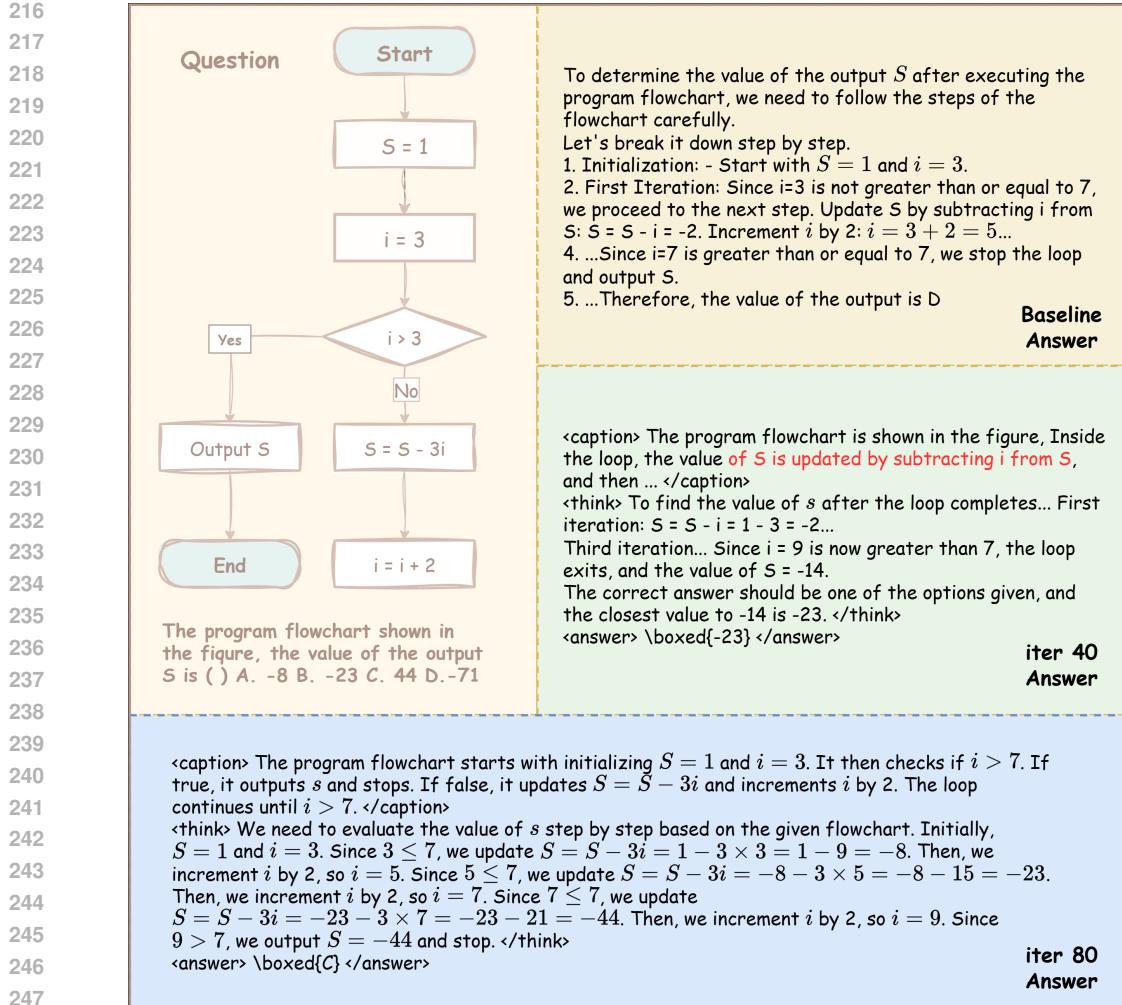


Figure 2: Evolution of caption quality during CaR training on a flowchart reasoning problem. **Baseline** (yellow): Direct reasoning without explicit caption produces incorrect answer (D). **Iter 40** (green): Early-stage caption captures partial information (“ S is updated by subtracting i from S ”) but misses initialization details, leading to approximate answer (-23). **Iter 80** (blue): Fully-trained model generates complete caption with precise variable initialization ($S = 1$, $i = 3$), loop condition ($i > 7$), and update rules ($S = S - 3i$, $i = i + 2$), enabling correct step-by-step reasoning and final answer (C: -44).

where $R_{\text{acc}} \in \{0, 1\}$ indicates final answer correctness against ground truth, $R_{\text{format}} \in \{0, 1\}$ ensures proper output structure (caption, reasoning, answer sections), and $R(C \mid I, Q)$ is the gain-based caption reward from Equation 1. We set weights $(w_{\text{acc}}, w_{\text{format}}, w_{\text{caption}}) = (1.0, 0.1, 1.0)$, giving equal importance to final accuracy and caption utility while using a lower format weight since format compliance is typically achieved early in training.

3.3 TRAINING IMPLEMENTATION

We integrate caption rewards with Group Relative Policy Optimization (GRPO) (30), which provides stable training without requiring a separate critic network.

For each training sample, we generate $n = 8$ caption candidates using temperature sampling ($\tau = 0.7$). Computing absolute rewards can lead to high variance across samples of different difficulty. Following GRPO, we normalize rewards within each group:

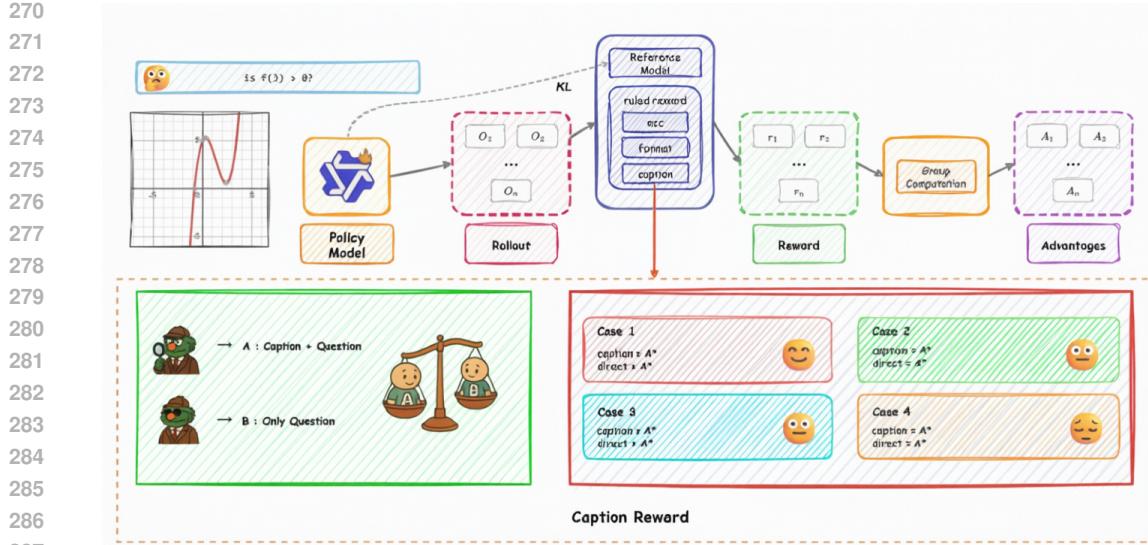


Figure 3: CaR training pipeline. For each image-question pair, the policy model generates multiple caption candidates. Each candidate is evaluated through both direct reasoning (without caption) and caption-enhanced reasoning to compute performance gains. The gain-based rewards are normalized within each group, and GRPO updates optimize the policy to favor captions with higher reasoning utility.

$$\hat{R}_i = \frac{R_i - \mu_g}{\sigma_g + \epsilon} \quad (3)$$

where μ_g and σ_g are the mean and standard deviation of rewards within the group, and $\epsilon = 10^{-8}$ prevents division by zero. This normalization ensures that the model receives meaningful gradient signals regardless of sample difficulty.

The policy is updated using clipped importance sampling to prevent large deviations from the reference policy:

$$\mathcal{L}(\theta) = -\mathbb{E}_{(I, Q, C) \sim \mathcal{D}} \left[\min \left(\rho_\theta \cdot \hat{R}, \text{clip}(\rho_\theta, 1 - \epsilon, 1 + \epsilon) \cdot \hat{R} \right) \right] \quad (4)$$

where $\rho_\theta = \pi_\theta(C \mid I, Q) / \pi_{\text{ref}}(C \mid I, Q)$ is the importance ratio between the current policy and reference policy, and $\epsilon = 0.2$ is the clipping threshold. The reference policy is updated periodically (every 100 steps) to track the evolving policy while maintaining training stability.

A critical design choice is how to assess answer correctness for computing R_{acc} and the gain-based reward. Using the training model itself would create a circular dependency where the model judges its own outputs. Instead, we employ an external evaluator—either GPT-4o-mini or Qwen2.5-7B-Instruct—to determine semantic equivalence between predicted and ground-truth answers.

The evaluator receives the question, ground-truth answer, and predicted answer, then outputs a binary correctness judgment. This design has two advantages: (1) it provides unbiased assessment independent of the training model's beliefs; (2) it handles semantic equivalence (e.g., “0.5” vs “1/2” vs “half”) better than exact string matching. We found that GPT-4o-mini provides more consistent judgments with lower variance across similar samples, so we use it as the default evaluator.

To ensure clean separation between perception and reasoning phases, we enforce a structured output format with explicit XML-style tags:

```
<caption>...</caption> <think>...</think> <answer>...</answer>
```

324
 325 Table 1: Comprehensive performance comparison across visual reasoning benchmarks. CaR denotes
 326 our Caption as Reward method.

	MMK12	MathVista	MathVision	MathVerse	DynaMath	WeMath	LogicVista	OlympiadBench	Avg
<i>Qwen2.5-VL-3B</i>									
GRPO	57.6	66.1	25.3	30.1	14.4	28.8	42.1	8.8	34.2
Ours	60.1	68.1	26.2	31.1	15.0	30.1	41.4	8.8	35.1
<i>InternVL2.5-4B</i>									
GRPO	58.8	67.5	26.0	30.8	14.9	29.4	43.4	8.6	34.9
Ours	61.6	69.7	27.0	31.9	15.4	30.3	43.5	8.6	36.0
<i>Qwen2.5-VL-7B</i>									
GRPO	61.2	71.5	27.3	32.5	15.5	31.0	45.8	9.0	36.7
Ours	63.8	74.3	28.5	33.9	16.2	32.3	46.4	9.2	38.1
<i>Baseline Comparisons (3B)</i>									
Qwen2.5-VL-3B	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

341
 342 The `<caption>` section contains the visual description, `<think>` contains the reasoning process,
 343 and `<answer>` contains the final response. This structure allows us to extract and evaluate each
 344 component independently during reward computation. Format compliance is checked automatically,
 345 and $R_{\text{format}} = 1$ only when all three sections are present and properly formatted.

346
 347 We sample 30K visual question-answer pairs from two existing datasets: MM-Eureka (17) (mathematical
 348 reasoning) and VirL39K (33) (logical reasoning). Importantly, we use only the question-answer
 349 pairs without any chain-of-thought annotations or human-written captions. The model must learn to
 350 generate useful captions and reasoning chains purely through reinforcement learning signals. This
 351 demonstrates that CaR can improve visual reasoning without expensive annotation efforts.

352 4 EXPERIMENTS

353 4.1 EXPERIMENTAL SETUP

354
 355 We train models using two established multimodal reasoning datasets: MM-Eureka (17), focusing
 356 on mathematical problem-solving, and VirL39K (33), targeting logical inference tasks. We apply
 357 stratified sampling based on difficulty levels determined through eight inference runs with Qwen2.5-
 358 VL-7B (34), creating a balanced training set of 30K samples with a 90%/10% mix of medium
 359 and high-difficulty items. For evaluation, we assess performance across eight challenging visual
 360 reasoning benchmarks: MathVista (23), MathVision, MathVerse, DynaMath, WeMath, LogicVista,
 361 MMK12-EVAL, and OlympiadBench. All experiments use Qwen2.5-VL models at 3B and 7B
 362 parameter scales with GRPO (group size $n = 8$, learning rate 5×10^{-7} , temperature 0.9) for two
 363 epochs. The composite reward weights are set to (1.0, 0.1, 1.0) for accuracy, format, and caption
 364 components respectively, with GPT-4o-mini or Qwen2.5-7B-Instruct serving as external evaluators.
 365 Training is conducted using the VeRL framework on NVIDIA A100 GPUs.

366 4.2 MAIN RESULTS

367
 368 Table 1 presents comprehensive results across all evaluation benchmarks. CaR demonstrates consistent
 369 improvements over baseline methods across different model scales. The 3B model achieves 35.1%
 370 average accuracy, improving +0.9 points over the GRPO baseline (34.2%) and +5.3 points over the
 371 3B-Instruct baseline (29.8%).

372
 373 Notably, CaR shows particularly strong improvements on individual benchmarks. On MMK12-EVAL,
 374 the 3B model improves from 57.6% to 60.1% (+2.5 points), and on MathVista from 66.1% to 68.1%
 375 (+2.0 points). The 7B model exhibits even larger gains, improving from 36.7% to 38.1% average
 376 (+1.4 points), with notable improvements on MathVista (71.5% to 74.3%, +2.8 points) and MMK12
 377 (61.2% to 63.8%, +2.6 points).

378

379
380
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				
	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	35.7			
Ours	21.9	28.2	60.8	52.2	0.0	33.1	37.2	78.4	71.5	26.9	41.0			

385

386

387

388
Table 3: Impact of different external evaluators on caption reward quality.

	MMK12	MathVista	MathVision	MathVerse	DynaMath	WeMath	LogicVista	OlympiadBench	Avg
Qwen2.5-7B-Instruct	50.4	65.1	22.0	30.5	12.8	25.2	38.9	7.5	31.6
gpt-4o-mini	49.8	66.3	22.0	31.6	12.3	27.2	43.4	7.9	32.8

393

394

395
To validate that CaR’s effectiveness extends beyond a single model family, we also evaluate on
396 InternVL2.5-4B (4). Results show consistent improvements (+1.1 points average), demon-
397 strating that the gain-based reward mechanism transfers across different VLM archi-
398 tectures. This cross-architecture consistency suggests that CaR captures fundamental properties of useful visual
399 descriptions rather than architecture-specific patterns.400
Performance improvements vary across benchmarks, with the largest gains on tasks requiring precise
401 visual extraction. MathVista shows consistent improvement across all model scales, while DynaMath
402 and OlympiadBench—which involve complex multi-step reasoning—show more modest gains. This
403 pattern suggests that CaR primarily improves the perception component of visual reasoning, with
404 downstream reasoning benefiting indirectly from more accurate visual information.405
Compared to recent competitive baselines, our 3B model (35.1%) outperforms Visionary-R1 (33.0%),
406 TBAC-VLR1 (33.8%), and VLAA-Thinker-3B (32.9%) on average accuracy across all benchmarks.
407

408

409

410
4.3 VISUAL PERCEPTION ANALYSIS

411

412
To evaluate CaR’s impact on visual perception specifically, we conduct additional experiments
413 on MME-RealWorld, a benchmark designed for fine-grained visual understanding across diverse
414 real-world scenarios. Table 2 shows detailed results across reasoning and perception tasks.415
CaR demonstrates substantial improvements in visual perception capabilities, achieving 41.0%
416 average performance compared to 35.7% for the baseline (+5.3 points). The most significant gains
417 occur in Perception-Diagram (+31.4 points: 71.5 vs 40.1) and Perception-OCR (+8.1 points: 78.4
418 vs 70.3). These results confirm that CaR’s gain-based reward mechanism effectively enhances the
419 model’s ability to extract task-relevant visual information.420
The performance gains are not uniform across all perception categories. Diagram understanding
421 shows the largest improvement, likely because mathematical diagrams require precise extraction of
422 numerical values, geometric relationships, and labels—exactly the type of information that CaR’s
423 gain-based reward incentivizes. OCR tasks also benefit substantially, as accurate text recognition
424 is often critical for correct reasoning. In contrast, autonomous driving and remote sensing tasks
425 show more modest improvements, possibly because these domains require different types of visual
426 understanding that are less directly captured by our question-answering training format.427
Interestingly, some reasoning subtasks show slight decreases (e.g., Monitor: 22.5 to 21.9, Auto Drive:
428 30.0 to 28.2) while corresponding perception tasks improve. This suggests that CaR may shift the
429 model’s focus toward detailed visual extraction at a small cost to high-level reasoning in certain
430 domains. However, the overall average improvement (+5.3 points) indicates that the perception gains
431 substantially outweigh any reasoning trade-offs, validating CaR’s design philosophy of prioritizing
accurate visual understanding as a foundation for reasoning.

432 4.4 ANALYSIS
433434 We investigate the impact of different evaluation models for caption reward calculation (Table 3). GPT-
435 40-mini yields superior performance (32.8% average) compared to Qwen2.5-7B-Instruct (31.6%),
436 with a +1.2 points difference. This suggests that more capable evaluators provide cleaner training
437 signals with reduced evaluation noise, highlighting the importance of evaluator quality in RL-based
438 training pipelines.439 Analysis of 100 samples where CaR succeeded but baselines failed reveals consistent improvement
440 patterns. First, CaR generates specific numerical values and spatial relationships rather than approxi-
441 mate descriptions—for example, describing “a triangle with sides 3, 4, and 5 units” instead of “a right
442 triangle.” Second, models learn to prioritize information directly relevant to the question, emphasizing
443 geometric relationships in math problems over aesthetic details like colors or backgrounds. Third, the
444 resulting descriptions follow structured patterns that systematically cover different image regions,
445 reducing the likelihood of missing critical visual elements.446 Regarding computational cost, CaR introduces additional inference passes for caption evaluation.
447 However, several optimizations minimize overhead: (1) the question-only inference results can be
448 precomputed before training, eliminating network I/O during the training loop; (2) during rollout,
449 partially generated captions that share common prefixes can reuse cached computations; (3) the
450 caption evaluation and importance sampling can be computed in parallel, synchronizing only at the
451 final loss computation to avoid blocking the training pipeline. With these optimizations, CaR remains
452 competitive with other RL-based methods in terms of training efficiency.453
454 5 DISCUSSION AND LIMITATIONS455 Our analysis suggests several factors contribute to CaR’s effectiveness. By directly optimizing for
456 task utility rather than proxy metrics like linguistic similarity, CaR creates a tight feedback loop
457 between perception quality and reasoning accuracy. The gain-based reward also naturally establishes
458 an implicit curriculum where the model learns to fix obvious perception errors before gradually
459 improving on subtler cases—this organic difficulty progression may explain why CaR scales better
460 with data than SFT. Additionally, separate evaluation of direct and caption-enhanced reasoning
461 provides disentangled learning signals that clarify what visual information is missing, helping the
462 model identify task-relevant visual features rather than memorizing caption patterns.463 Unlike linguistic metrics (BLEU, ROUGE) that measure surface-level similarity, CaR’s gain-based
464 reward directly captures reasoning utility. This fundamental difference explains why CaR outperforms
465 caption-based methods that rely on linguistic optimization. Compared to self-consistency approaches
466 that check reasoning agreement, CaR provides stronger supervision by requiring captions to demon-
467 strably improve task performance. The use of external evaluators also avoids the self-reinforcement
468 bias that can occur when models judge their own outputs.469 Our experiments demonstrate that CaR generalizes across different model scales (3B, 4B, 7B) and task
470 types (mathematical reasoning, logical inference, visual perception). The consistent improvements
471 suggest that the gain-based reward captures fundamental aspects of visual description quality that
472 transfer across domains. However, the degree of improvement varies by task: CaR shows larger
473 gains on tasks requiring precise numerical extraction (e.g., diagram understanding) compared to tasks
474 dominated by spatial reasoning.475 While CaR demonstrates strong results, several limitations warrant discussion. The approximately $3 \times$
476 computational overhead compared to SFT may limit adoption for resource-constrained settings; future
477 work should explore more efficient reward computation strategies such as caching evaluator responses
478 or using lightweight reward models. We evaluated CaR only on Qwen2.5-VL (34) and InternVL (4)
479 models, and testing on more diverse architectures (LLaVA (20; 21), BLIP (15), MiniGPT-4 (40))
480 would strengthen claims about generalizability. Current experiments focus on question-answering
481 tasks, and extending CaR to other modalities (video, audio) and tasks (generation, editing) remains
482 unexplored. Our reward weights and thresholds were determined through limited grid search, and
483 more principled approaches using multi-objective optimization or learned reward functions could
484 improve performance.

486
487
488
489
490
491
492
493
494
495
496
497
498
499
500
501
502
503
504
505
506
507
508
509
510
511
512
513
514
515
516
517
518
519
520
521
522
523
524
525
526
527
528
529
530
531
532
533
534
535
536
537
538
539
CaR’s improved visual perception could enable more reliable VLM deployments in education, accessibility, and scientific research. For educational applications, accurate visual understanding is critical for tutoring systems that explain diagrams and figures. In accessibility contexts, better caption generation could improve screen reader experiences for visually impaired users. 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 to detect potential misuse.

6 CONCLUSION

We introduced Caption as Reward (CaR), a reinforcement learning framework that optimizes visual descriptions based on their utility for downstream reasoning tasks. CaR addresses a fundamental limitation in vision-language model training: the disconnect between linguistic caption quality and reasoning performance. By evaluating captions through performance gains rather than linguistic similarity metrics, CaR provides a principled training signal that directly aligns visual understanding with task requirements.

Our experimental results demonstrate CaR’s effectiveness across eight challenging visual reasoning benchmarks. The 3B model achieves 35.1% average accuracy, outperforming the instruction-tuned baseline (29.8%, +5.3 points) and the GRPO baseline (34.2%, +0.9 points). The 7B model shows consistent improvements from 36.7% to 38.1% (+1.4 points), demonstrating scalability across model sizes. Cross-architecture validation on InternVL2.5-4B (+1.1 points) confirms that CaR’s benefits transfer beyond a single model family.

The key insight underlying CaR is that visual description quality should be measured by reasoning utility rather than descriptive fluency. This performance-centric approach enables models to learn what visual information matters for specific tasks, resulting in more precise and task-relevant visual understanding. Additional evaluation on MME-RealWorld confirms enhanced perception capabilities, with particularly strong improvements in diagram understanding (+31.4 points) and OCR tasks (+8.1 points).

Current limitations include evaluation on two model families (Qwen2.5-VL and InternVL) and reliance on external evaluators for reward computation. Future work will explore broader architectural validation across diverse VLM designs, extension to video and audio modalities, and application to generative tasks beyond question answering.

CaR opens new directions for task-adaptive multimodal learning by demonstrating that performance-based optimization can effectively bridge the perception-reasoning gap in vision-language models without requiring human annotations or auxiliary reward models. This approach may generalize to other domains where proxy metrics (e.g., perplexity, BLEU) poorly correlate with downstream task performance, suggesting a broader paradigm of utility-based optimization for foundation models.

540
541 ETHICS STATEMENT

542 CaR targets educational and scientific reasoning workloads by strengthening factual visual under-
 543 standing. Although stronger perception could be misused to generate misleading analyses, the method
 544 reduces hallucinated descriptions and relies only on public datasets, which we acknowledge may
 545 carry existing societal biases. We adhere to the ICLR Code of Ethics and confirm that our work
 546 complies with all ethical guidelines. All authors of this work have read and commit to adhering to the
 547 ICLR Code of Ethics.

548
549 REPRODUCIBILITY STATEMENT
550

551 We train on the publicly available MM-Eureka and VirL39K corpora using Qwen2.5-VL models
 552 (3B/7B) and InternVL2.5-4B. All hyperparameters (GRPO with $n=8$, learning rate 5×10^{-7} ,
 553 temperature 0.9, two epochs, reward weights (1.0, 0.1, 1.0)) and evaluator choices (GPT-4o-mini
 554 or Qwen2.5-7B-Instruct) are described in Section 4.1. Code, data splits, and checkpoints will be
 555 released upon acceptance.

556
557 REFERENCES
558

- [1] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. *arXiv preprint arXiv:2308.12966*, 2023. URL <https://arxiv.org/abs/2308.12966>.
- [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] Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Conghui He, Jiaqi Wang, Feng Zhao, and Dahua Lin. ShareGPT4V: Improving large multi-modal models with better captions. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2024. URL <https://arxiv.org/abs/2311.12793>.
- [4] Zhe Chen, Jiannan Wu, Wenhui Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qingsong Zhang, Xizhou Zhu, Lewei Lu, Bin Li, Ping Luo, Tong Lu, Yu Qiao, and Jifeng Dai. InternVL: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 24185–24198, 2024. URL <https://arxiv.org/abs/2312.14238>.
- [5] Yihe Deng, Hritik Bansal, Fan Yin, Nanyun Peng, Wei Wang, and Kai-Wei Chang. Open-vlthinker: Complex vision-language reasoning via iterative SFT-RL cycles. *arXiv preprint arXiv:2503.17352*, 2025. URL <https://arxiv.org/abs/2503.17352>.
- [6] 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>.
- [7] Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, Yunsheng Wu, and Rongrong Ji. MME: A comprehensive evaluation benchmark for multimodal large language models. *arXiv preprint arXiv:2306.13394*, 2024. URL <https://arxiv.org/abs/2306.13394>.
- [8] Adam Gleave and Geoffrey Irving. Uncertainty estimation for language reward models. *arXiv preprint arXiv:2203.07472*, 2022. URL <https://arxiv.org/abs/2203.07472>.
- [9] 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. *Nature*, 2025.

594 doi: 10.1038/s41586-025-09422-z. URL <https://www.nature.com/articles/s41586-025-09422-z>.

595

596 [10] Huanjin Hu, Dahu Tu, Siyuan Wei, Hongwei Gao, Hui Zhang, Xiaolong Li, and Ming-Hsuan Yang. Dense connector for MLLMs. *arXiv preprint arXiv:2405.13800*, 2024. URL <https://arxiv.org/abs/2405.13800>.

597

598 [11] Wenxuan Huang, Bohan Shen, Zihao Zhu, Haoyang Zhang, and Lei Wang. Vision-r1: Incentivizing reasoning capability in multimodal large language models. *arXiv preprint arXiv:2503.06749*, 2025. URL <https://arxiv.org/abs/2503.06749>.

599

600 [12] 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>.

601

602 [13] Harrison Lee, Samrat Phatale, Hassan Mansoor, Thomas Mesnard, Johan Ferret, Kellie Lu, Colton Bishop, Ethan Hall, Victor Carbune, Abhinav Rastogi, and Sushant Prakash. RLAIF vs. RLHF: Scaling reinforcement learning from human feedback with AI feedback. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2024. URL <https://arxiv.org/abs/2309.00267>.

603

604 [14] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. BLIP: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *Proceedings of the International Conference on Machine Learning (ICML)*, volume 162, pp. 12888–12900, 2022. URL <https://arxiv.org/abs/2201.12086>.

605

606 [15] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *Proceedings of the International Conference on Machine Learning (ICML)*, volume 202, pp. 20351–20383, 2023. URL <https://arxiv.org/abs/2301.12597>.

607

608 [16] Yunxin Li, Zhenyu Liu, Zitao Li, Xuanyu Zhang, Zhenran Xu, Xinyu Chen, Haoyuan Shi, Shenyuan Jiang, Xintong Wang, Jifang Wang, Shouzheng Huang, Xinping Zhao, Borui Jiang, Lanqing Hong, Longyue Wang, Zhuotao Tian, Baoxing Huai, Wenhan Luo, Weihua Luo, Zheng Zhang, Baotian Hu, and Min Zhang. Perception, reason, think, and plan: A survey on large multimodal reasoning models. *arXiv preprint arXiv:2505.04921*, 2025. URL <https://arxiv.org/abs/2505.04921>.

609

610 [17] Zhijie Li, Hao Wang, Yu Chen, and Ming Zhang. Mm-eureka: A multi-modal reasoning dataset for mathematical problem solving. *arXiv preprint arXiv:2405.14677*, 2024. URL <https://arxiv.org/abs/2405.14677>.

611

612 [18] 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>.

613

614 [19] Hunter Lightman, Vineet Kosaraju, Yura Burda, Harri Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's verify step by step. *arXiv preprint arXiv:2305.20050*, 2023. URL <https://arxiv.org/abs/2305.20050>.

615

616 [20] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *Advances in Neural Information Processing Systems*, volume 36, pp. 34892–34916, 2023. URL <https://arxiv.org/abs/2304.08485>.

617

618 [21] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 26296–26306, 2024. URL <https://arxiv.org/abs/2310.03744>.

619

620 [22] Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, Kai Chen, and Dahu Lin. MMBench: Is your multi-modal model an all-around player? In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2024. URL <https://arxiv.org/abs/2307.06281>.

621

622

623

624

625

626

627

628

629

630

631

632

633

634

635

636

637

638

639

640

641

642

643

644

645

646

647

648 [23] Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng,
 649 Kai-Wei Chang, Michel Galley, and Jianfeng Gao. MathVista: Evaluating mathematical reason-
 650 ing of foundation models in visual contexts. In *Proceedings of the International Conference on*
 651 *Learning Representations (ICLR)*, 2024. URL <https://arxiv.org/abs/2310.02255>.
 652

653 [24] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin,
 654 Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton,
 655 Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F Christiano,
 656 Jan Leike, and Ryan Lowe. Training language models to follow instructions with human
 657 feedback. In *Advances in Neural Information Processing Systems*, volume 35, pp. 27730–27744,
 658 2022. URL <https://arxiv.org/abs/2203.02155>.
 659

660 [25] Yi Peng, Peiyu Wang, Xiaokun Wang, Yichen Wei, Jiangbo Pei, Weijie Qiu, Ai Jian, Yunzhuo
 661 Hao, Jiachun Pan, Tianyidan Xie, Li Ge, Rongxian Zhuang, Xuchen Song, Yang Liu, and Yahui
 662 Zhou. Skywork r1v: Pioneering multimodal reasoning with chain-of-thought. *arXiv preprint*
 663 *arXiv:2504.05599*, 2025. URL <https://arxiv.org/abs/2504.05599>.
 664

665 [26] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agar-
 666 wal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya
 667 Sutskever. Learning transferable visual models from natural language supervision. In *Proceed-
 668 ings of the International Conference on Machine Learning (ICML)*, volume 139, pp. 8748–8763,
 669 2021. URL <https://arxiv.org/abs/2103.00020>.
 670

671 [27] Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and
 672 Chelsea Finn. Direct preference optimization: Your language model is secretly a reward
 673 model. In *Advances in Neural Information Processing Systems*, volume 36, 2023. URL
 674 <https://arxiv.org/abs/2305.18290>.
 675

676 [28] Juan Rocamonde, Victoriano Montesinos, Elvis Nava, Ethan Perez, and David Lindner. Vision-
 677 language models are zero-shot reward models for reinforcement learning. In *Proceedings of the*
 678 *International Conference on Learning Representations*, 2024. URL <https://arxiv.org/abs/2310.12921>.
 679

680 [29] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal
 681 policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017. URL <https://arxiv.org/abs/1707.06347>.
 682

683 [30] Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Mingchuan Zhang, Y. K. Li,
 684 Y. Wu, and Daya Guo. Deepseekmath: Pushing the limits of mathematical reasoning in open
 685 language models. *arXiv preprint arXiv:2402.03300*, 2024. URL <https://arxiv.org/abs/2402.03300>. Contains GRPO (Group Relative Policy Optimization) method.
 686

687 [31] Joar Skalse, Nikolaus Howe, Dmitrii Krasheninnikov, and David Krueger. Defining and charac-
 688 terizing reward hacking. In *Advances in Neural Information Processing Systems*, volume 35,
 689 pp. 16748–16761, 2022. URL <https://arxiv.org/abs/2209.13085>.
 690

691 [32] Haozhe Wang, Chao Qu, Zuming Huang, Wei Chu, Fangzhen Lin, and Wenhui Chen. VI-
 692 rethinker: Incentivizing self-reflection of vision-language models with reinforcement learning.
 693 In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025. URL <https://arxiv.org/abs/2504.08837>.
 694

695 [33] Lei Wang, Xiaoming Liu, Yifan Zhang, and Hao Chen. Virl39k: A large-scale visual reasoning
 696 dataset for logical inference. *arXiv preprint arXiv:2404.12890*, 2024. URL <https://arxiv.org/abs/2404.12890>.
 697

698 [34] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing
 699 Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men,
 700 Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. Qwen2-vl: Enhancing vision-
 701 language model’s perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*,
 702 2024. URL <https://arxiv.org/abs/2409.12191>.
 703

702 [35] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi,
 703 Quoc V Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language
 704 models. In *Advances in Neural Information Processing Systems*, volume 35, pp. 24824–24837,
 705 2022. URL <https://arxiv.org/abs/2201.11903>.

706 [36] Guowei Xu, Peng Liu, Changyong Zhu, Wenhui Lin, Haotian Shao, Wei Zhang, Mingyu Gao,
 707 and Kai Zhang. LLaVA-CoT: Let vision language models reason step-by-step. *arXiv preprint*
 708 *arXiv:2411.10440*, 2024. URL <https://arxiv.org/abs/2411.10440>.

710 [37] Yufei Zhan, Yousong Zhu, Shurong Zheng, Hongyin Zhao, Fan Yang, Ming Tang, and Jinqiao
 711 Wang. Vision-r1: Evolving human-free alignment in large vision-language models via vision-
 712 guided reinforcement learning. *arXiv preprint arXiv:2503.18013*, 2025. URL <https://arxiv.org/abs/2503.18013>.

714 [38] Shengjia Zhang, Junjie Wu, Jiawei Chen, Changwang Zhang, Xingyu Lou, Wangchunshu
 715 Zhou, Sheng Zhou, Can Wang, and Jun Wang. Othink-r1: Intrinsic fast/slow thinking mode
 716 switching for over-reasoning mitigation. *arXiv preprint arXiv:2506.02397*, 2025. URL <https://arxiv.org/abs/2506.02397>.

719 [39] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang,
 720 Zi Lin, Zhuohan Li, Dacheng Li, Eric P Xing, Hao Zhang, Joseph E Gonzalez, and Ion Stoica.
 721 Judging LLM-as-a-judge with mt-bench and chatbot arena. In *Advances in Neural Information
 722 Processing Systems*, volume 36, pp. 46595–46623, 2023. URL <https://arxiv.org/abs/2306.05685>.

724 [40] Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. MiniGPT-4:
 725 Enhancing vision-language understanding with advanced large language models. In *Proceedings
 726 of the International Conference on Learning Representations (ICLR)*, 2024. URL <https://arxiv.org/abs/2304.10592>.

729

730 A TECHNICAL APPENDICES AND SUPPLEMENTARY MATERIAL

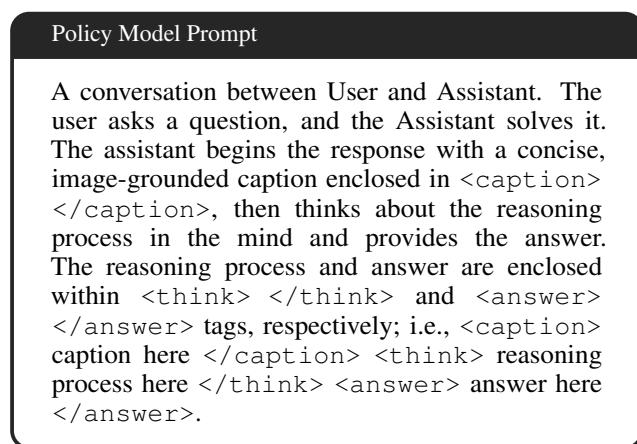
731

732 A.1 POLICY MODEL PROMPT

733

734 To ensure the model generates structured visual descriptions before reasoning, we include additional
 735 instructions in the system prompt to guide the policy model in generating the corresponding output.
 736 Using this prompt, the model will insert the corresponding image description labeled as $\langle\text{caption}_i\rangle$
 737 before the thinking process, additional to the existing $\langle\text{think}_i\rangle$ and $\langle\text{answer}_i\rangle$.

738



754 A.2 GAIN-BASED REWARD IMPLEMENTATION

755

Algorithm 1 summarizes the reward computation pipeline referenced in Section 3.2.

756 **Algorithm 1** Gain-Based Reward Computation

757 **Require:** Image I , Question Q , Generated Caption C , Ground Truth A^* , Vision-Language Model
 758 \mathcal{M} , External Evaluator \mathcal{E}

759 **Ensure:** Gain-based reward $R(C|I, Q)$

760 1: $A_{\text{direct}} \leftarrow \mathcal{M}(I, Q)$ {Direct reasoning}

761 2: $A_{\text{caption}} \leftarrow \mathcal{M}(I, Q, C)$ {Caption-enhanced reasoning}

762 3: $\text{correct}_{\text{direct}} \leftarrow \mathcal{E}.\text{SemanticMatch}(A_{\text{direct}}, A^*)$

763 4: $\text{correct}_{\text{caption}} \leftarrow \mathcal{E}.\text{SemanticMatch}(A_{\text{caption}}, A^*)$

764 5: **if** $\text{correct}_{\text{caption}}$ **and** $\neg\text{correct}_{\text{direct}}$ **then**

765 6: $R(C|I, Q) \leftarrow 1.0$ {Caption fixes an error}

766 7: **else if** $\text{correct}_{\text{caption}}$ **and** $\text{correct}_{\text{direct}}$ **then**

767 8: $R(C|I, Q) \leftarrow 0.7$ {Caption confirms success}

768 9: **else if** $\neg\text{correct}_{\text{caption}}$ **and** $\neg\text{correct}_{\text{direct}}$ **then**

769 10: $R(C|I, Q) \leftarrow 0.2$ {Both attempts fail}

770 11: **else**

771 12: $R(C|I, Q) \leftarrow 0.0$ {Caption harms accuracy}

772 13: **end if**

773 14: **return** $R(C|I, Q)$

774
775 A.3 TRAINING DATA SOURCES
776

777 Our training dataset combines two established open-source multimodal reasoning corpora: MM-
 778 Eureka (17) and VirL39K (33), covering diverse visual reasoning tasks including mathematical
 779 problem-solving and spatial reasoning. We apply stratified sampling based on difficulty levels to
 780 ensure effective training data quality, following the data construction methodology described in
 781 Section 3.3.

782
783 A.4 ADDITIONAL EXPERIMENTAL RESULTS
784

785 **Data Scaling Analysis** We analyze the impact of training data size on CaR performance. Table 4
 786 shows results with 10K, 20K, and 30K training samples on the 3B model. CaR demonstrates
 787 consistent improvements over GRPO across all data scales: 10K (+0.6 points), 20K (+1.2 points),
 788 and 30K (+1.3 points). This scaling trend indicates that CaR effectively leverages additional training
 789 data to enhance visual reasoning capabilities.

790
791 Table 4: Impact of training data size on CaR performance (Qwen2.5-VL-3B).

Data Size		Method	MMK12	MathVista	MathVision	MathVerse	DynaMath	WeMath	LogicVista	OlympiadBench	Avg
10K	SFT	29.6	54.7	18.8	8.1	9.7	16.2	34.9	6.0	22.2	
10K	GRPO	48.5	64.1	22.6	32.0	12.5	27.0	41.2	8.0	32.0	
10K	CaR	49.8	66.3	22.0	31.6	12.3	27.7	43.4	7.9	32.6	
20K	SFT	33.7	54.0	18.6	9.1	9.2	18.2	34.0	6.0	22.9	
20K	GRPO	51.2	64.8	23.4	28.5	13.1	27.5	41.8	8.2	32.3	
20K	CaR	53.5	66.9	24.2	29.4	13.6	28.4	43.1	8.5	33.5	
30K	SFT	38.5	61.5	19.1	23.6	4.3	17.9	36.7	7.1	26.1	
30K	GRPO	54.8	63.9	24.5	29.2	13.9	27.9	40.8	8.5	32.9	
30K	CaR	57.6	66.1	25.3	30.1	14.4	28.8	42.1	8.8	34.2	

802
803 **Hyper-parameter Analysis** We analyze the sensitivity of CaR to the caption reward weight
 804 w_{caption} in the composite reward structure. For the accuracy and format rewards, we follow standard
 805 GRPO settings with equal weights. Table 5 shows results across all benchmarks for different caption
 806 reward weights. We observe that $w_{\text{caption}} = 0.1$ achieves the best overall performance (35.1%
 807 average), while higher weights lead to slight degradation. To maintain training stability and prevent
 808 caption reward from dominating the optimization objective, we recommend using $w_{\text{caption}} = 0.1$ in
 809 practice.

810

811

Table 5: Hyper-parameter analysis for caption reward weight $w_{caption}$ (Qwen2.5-VL-3B).

812

813

814

815

Weight $w_{caption}$	MMK12	MathVista	MathVision	MathVerse	DynaMath	WeMath	LogicVista	OlympiadBench	Avg
0.1	60.1	68.1	26.2	31.1	15.0	30.1	41.4	8.8	35.1
0.5	57.0	64.7	24.6	29.5	14.1	28.3	39.1	8.3	33.2
1.0	55.3	62.9	24.2	28.7	13.8	27.7	38.3	8.1	32.4

816

817

818

819

820

A.5 VISUALIZATION OF THE CaR OUTPUT

821

822

823

We present four representative examples demonstrating CaR’s structured output format with caption, thinking, and answer components.

824

825

826

827

828

829

830

831

832

833

834

835

836

837

838

839

840

841

842

843

844

845

846

847

848

849

850

851

852

853

854

855

856

857

858

859

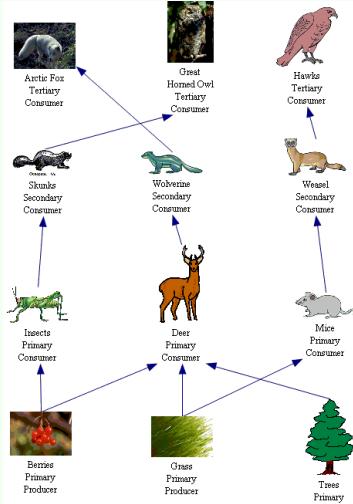
860

861

862

863

Diagram Reasoning



CaR Output

<caption> A food web diagram showing ecological relationships. Primary producers (berries, grass, trees) are at the bottom. Primary consumers include insects, deer, and mice. Secondary consumers are skunks, wolverine, and weasel. Tertiary consumers at top are arctic fox, great horned owl, and hawks. Blue arrows show energy flow from prey to predator. **</caption>**

<think> Looking at the food web, I need to find which animals eat mice. Following the arrows FROM mice, I can see only one arrow pointing to weasel. The weasel is the only predator that consumes mice in this diagram. **</think>**

<answer> 1 **</answer>**

Q: How many predators does the mouse have?

Options: 4, 1, 0, 2

A: 1

Table Understanding

Refrigerator	Initial temperature (°C)	Final temperature (°C)
Refrigerator with its door open	3.9	11.4
Refrigerator with its door closed	3.9	5.4

Q: During this time, thermal energy was transferred from () to ().

- A. each refrigerator...the surroundings
- B. the surroundings...each refrigerator

A: B

CaR Output

<caption> A table showing temperature changes in two refrigerators. Row 1: Refrigerator with door open, initial temp 3.9°C, final temp 11.4°C. Row 2: Refrigerator with door closed, initial temp 3.9°C, final temp 5.4°C. Both refrigerators increased in temperature over 10 minutes. **</caption>**

<think> Both refrigerators started at 3.9°C and ended at higher temperatures (11.4°C and 5.4°C). Since temperature increased, thermal energy must have flowed INTO the refrigerators FROM the warmer surroundings, not the other way around. **</think>**

<answer> B **</answer>**

864

865

866

867

868

869

870

871

872

873

874

875

876

877

878

879

880

881

882

883

884

885

886

887

888

889

890

891

892

893

894

895

896

897

898

899

900

901

902

903

904

905

906

907

908

909

910

911

912

913

914

915

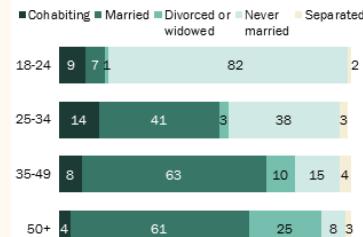
916

917

Chart Analysis

Adults ages 25 to 34 are most likely to be cohabiting

% who are _____, by age group



Note: All individuals who were cohabiting are included only in the cohabiting category, regardless of marital status. "Separated" includes those who were married, but their spouse was absent.

Source: Pew Research Center analysis of 2016 Current Population Survey, Annual Social and Economic Supplement (IPUMS).

PEW RESEARCH CENTER

Q: How many colors are in the bar?

A: 5

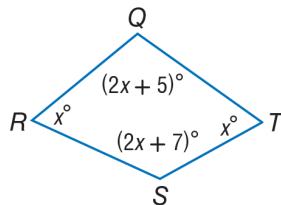
CaR Output

<caption> A horizontal stacked bar chart titled "Adults ages 25 to 34 are most likely to be cohabiting." Shows percentages by age group (18-24, 25-34, 35-49, 50+) with 5 categories: Cohabiting (dark green), Married (light green), Divorced/widowed (beige), Never married (olive), and Separated (gray). **</caption>**

<think> Counting the distinct colors in the legend: 1) Cohabiting (dark green), 2) Married (light green), 3) Divorced or widowed (beige), 4) Never married (olive/brown), 5) Separated (gray). Total: 5 colors. **</think>**

<answer> 5 **</answer>**

Geometry Problem



Q: Find $m\angle S$.

A: 123°

CaR Output

<caption> A quadrilateral QRST (kite shape) with vertices Q at top, R at left, S at bottom, T at right. Angle at R is x° , angle at T is x° . Angle at Q is $(2x+5)^\circ$, angle at S is $(2x+7)^\circ$. The shape appears symmetric about the vertical axis. **</caption>**

<think> Sum of angles in a quadrilateral = 360° . So: $x + x + (2x+5) + (2x+7) = 360$. Simplifying: $6x + 12 = 360$, thus $6x = 348$, $x = 58$. Therefore, $m\angle S = 2(58) + 7 = 116 + 7 = 123^\circ$. **</think>**

<answer> 123° **</answer>**