VERA: VISUAL EVIDENCE REINFORCEMENT FOR MULTIMODAL REASONING

Anonymous authors
Paper under double-blind review

ABSTRACT

Multimodal reasoning large language models (MRLMs) have made substantial progress, yet their reinforcement learning objectives still rely primarily on final-answer correctness, leaving intermediate visual evidence largely unsupervised. Such outcome-level feedback cannot determine whether a response is grounded in the image, preserves the information needed to solve the task, or reaches a correct answer through unreliable evidence. We introduce VERA, a visual evidence reinforcement framework that externalizes task-oriented visual evidence as a structured description separated from the reasoning process and final answer. VERA evaluates this evidence along two complementary dimensions: visual faithfulness, which requires every visual claim to be supported by the input image, and task sufficiency, which requires the question and evidence alone to recover the target answer. The two evidence rewards are combined with answer correctness and format compliance in group-relative policy optimization, allowing trajectories with similar answer outcomes to receive different credit according to their visual-faithfulness and task-sufficiency assessments. Across five MRLM backbones and eight multimodal reasoning benchmarks, VERA improves average accuracy by 5.0 percentage points while reducing output-token consumption by 9.8%–48.0%.

1 INTRODUCTION

Multimodal large language models have achieved substantial progress in visual question answering, chart understanding, and mathematical reasoning over visual inputs (Masry et al., 2022; Lu et al., 2024; Wang et al., 2024; Zhang et al., 2024). Recent reinforcement learning methods further improve the ability of multimodal reasoning large language models to perform complex multi-step reasoning (Huang et al., 2025b; Wei et al., 2025; Chen et al., 2025a). However, reliable multimodal reasoning requires more than producing the correct final answer. A model must first identify the visual information needed by the task and then reason correctly from that information. These two stages can fail independently, yet most existing reinforcement learning objectives collapse them into a single outcome-level reward based on final-answer correctness.

We begin by examining whether final-answer supervision adequately reflects the reliability of visual reasoning. The evidence suggests otherwise. Figure 1(a) shows that failures across visual reasoning benchmarks arise from both incorrect visual evidence and errors in subsequent reasoning. Models may misread an object, value, label, or spatial relationship, or they may correctly identify the relevant visual information but apply an invalid mathematical or logical inference. A final-answer reward observes only the resulting success or failure and therefore cannot identify which stage produced the error. Conversely, a correct answer does not establish that the intermediate visual interpretation was reliable, because the model may exploit textual regularities, make compensating errors, or arrive at the answer despite unsupported visual claims (Xia et al., 2025; Dong et al., 2025; Pan et al., 2026).

Longer reasoning trajectories do not by themselves resolve this supervision gap. Figure 1(b) shows that reasoning-oriented models produce substantially more output tokens than their instruction-oriented counterparts. Output length alone, however, does not assess whether the underlying visual evidence is faithful or sufficient. This observation motivates direct supervision of intermediate visual evidence rather than treating longer generation as a proxy for reliable reasoning.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

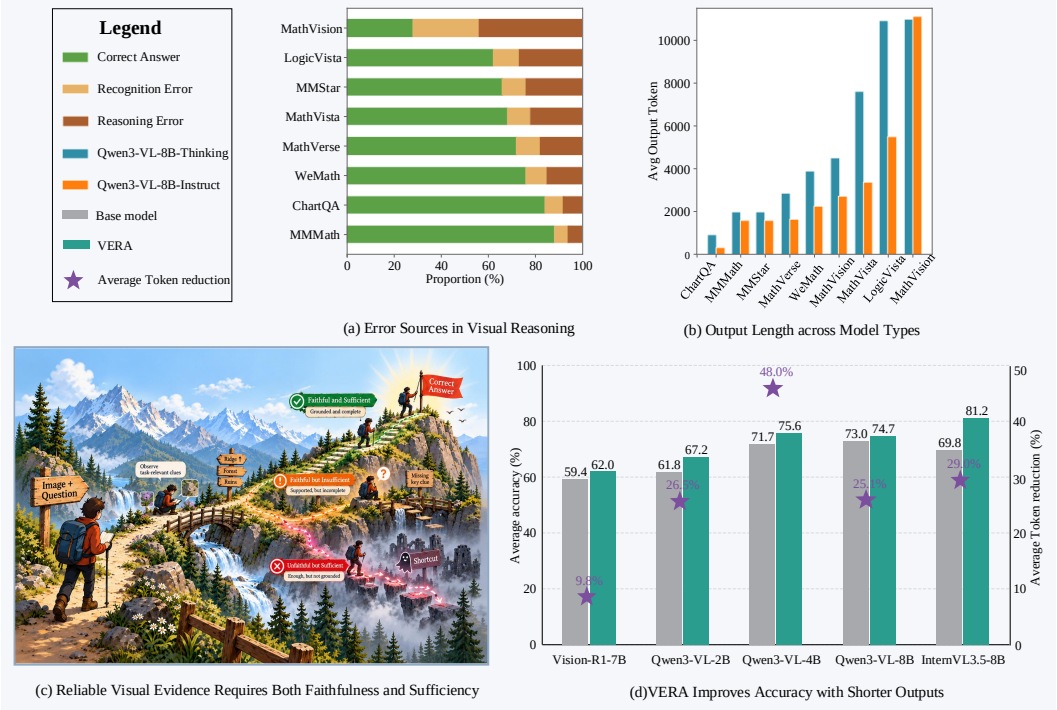


Figure 1: Motivation and overall effectiveness of VERA. (a) Distribution of correct responses, visual-recognition errors, and subsequent reasoning errors across visual reasoning benchmarks, showing that final-answer correctness alone cannot identify failures in intermediate visual evidence. (b) Comparison of output-token consumption between reasoning-oriented and instruction-oriented models, showing that reasoning-oriented models produce substantially longer outputs. Output length alone does not indicate whether the underlying visual evidence is faithful or sufficient. (c) Conceptual illustration of reliable task oriented visual evidence, which must be both faithful to the input image and sufficient for answering the target question. (d) Average accuracy of five multimodal reasoning backbones before and after applying VERA, together with the corresponding reductions in output-token consumption. VERA improves accuracy across all evaluated backbones while producing shorter outputs.

This observation raises a more fundamental question: what properties should intermediate visual evidence satisfy? As illustrated in Figure 1(c), we characterize reliable visual evidence through two complementary dimensions. *Visual faithfulness* requires every visual claim to be directly supported by the input image, whereas *task sufficiency* requires the evidence to preserve enough question-relevant information to determine the target answer. These properties are related but logically distinct. A description may be faithful while omitting a critical number, label, or spatial relation, making it insufficient for the task. Conversely, a description may contain enough information to recover the answer while also introducing unsupported visual claims. Supervising either dimension alone therefore leaves an important failure mode unresolved.

To address this problem, we introduce **VERA**, a visual evidence reinforcement framework for multimodal reasoning. VERA evaluates the generated evidence with two judges that receive different inputs and perform different roles. The visual-faithfulness judge receives the image and evidence description and determines whether every substantive visual claim is supported by the image. The task-sufficiency judge receives only the question and evidence description, without access to the original image, and attempts to recover the answer from the stated evidence. Its prediction is subsequently compared with the reference answer outside the judge. This input separation operationalizes the distinction between whether the evidence is visually grounded and whether it contains enough information to solve the task. The resulting visual-faithfulness and task-sufficiency rewards are combined with final-answer correctness and format compliance. For multiple responses sampled from the same input, VERA converts the aggregated rewards into group-relative advantages and op-

timizes the policy using DAPO (Yu et al., 2025). Consequently, trajectories with the same answer outcome can receive different optimization signals according to the quality of their intermediate visual evidence.

We evaluate VERA on five multimodal reasoning backbones with different architectures and scales across eight benchmarks covering general multimodal understanding, chart reasoning, visual logic, and visual mathematical reasoning. As summarized in Figure 1(d), VERA improves the eight-benchmark average accuracy for every evaluated backbone while reducing output-token consumption. The mean gain across the five backbone-level averages is 5.0 percentage points. For example, VERA improves Qwen3-VL-4B from 71.7% to 75.6% and InternVL3.5-8B from 69.8% to 81.2%. Meanwhile, output-token consumption is reduced by 9.8%–48.0% across the five backbones.

Our main contributions are summarized as follows:

- We identify a supervision gap in multimodal reinforcement learning: final-answer rewards do not distinguish failures in visual evidence acquisition from errors in subsequent reasoning. We characterize reliable task-oriented visual evidence through two complementary properties, visual faithfulness and task sufficiency.
- We propose VERA, which externalizes task-relevant visual evidence as a structured description separated from the reasoning process and final answer. VERA uses two input-separated judges to evaluate visual faithfulness and task sufficiency and incorporates the resulting evidence rewards into group-relative policy optimization.
- We evaluate VERA across five multimodal reasoning backbones and eight visual reasoning benchmarks. VERA consistently improves average accuracy while reducing output-token consumption, and controlled analyses characterize the effects of its evidence rewards and judge behavior.

2 Related Work

Multimodal Reasoning with Reinforcement Learning. Vision-R1 Huang et al. (2025b) highlights instability in direct reinforcement learning, motivating structured initialization Wei et al. (2025); Chen et al. (2025a). URSA Luo et al. (2025b) provides large-scale multimodal reasoning and process-supervision data, while R1-ShareVL Yao et al. (2025) shares trajectories and advantages to alleviate sparse rewards. Mixed-R1 Xu et al. (2025b) unifies rewards across heterogeneous tasks, and SophiaVL-R1 Fan et al. (2025a) combines outcome and process rewards with confidence-aware weighting. Other work improves sampling and optimization stability Wang et al. (2025c); Xu et al. (2025a).

Visual Evidence Grounding and Faithful Reasoning. GRIT Fan et al. (2025b) couples reasoning with bounding-box predictions, while v1 Chung et al. (2025) and Qwen Look Again Chu et al. (2025) encourage revisiting visual regions or tokens. Visionary-R1 Xia et al. (2025) and Ground What You See Pan et al. (2026) address visual shortcuts and insufficient grounding. From Sight to Insight Sharif et al. (2026) rewards visual information utilization. When Thinking Drifts Luo et al. (2025a) introduces Visual Evidence Reward for video reasoning, using a language-model judge to assess whether reasoning traces reference offline, teacher-generated visual facts. Point-RFT Ni et al. (2025) teaches reasoning interleaved with visual point coordinates through supervised fine-tuning, followed by reinforcement fine-tuning with format and answer-accuracy rewards.

Adaptive Tool Use and Reward Optimization. VL-Rethinker Wang et al. (2025b) and SRPO Wan et al. (2025) train reflection behaviors, while R-4B Yang et al. (2025) learns when to activate additional reasoning. DeepEyes Zheng et al. (2026) and Thyme Zhang et al. (2025) incorporate visual grounding and image operations into reasoning. Learning Only with Images Chen et al. (2025b) uses visual feedback for training, whereas AdaTooler-V Wang et al. (2025a) learns adaptive visual-tool invocation. Complementary optimization approaches, including Hint-GRPO Huang et al. (2025a) and DIVA-GRPO Gao et al. (2026), improve the use of training signals in multimodal reinforcement learning.

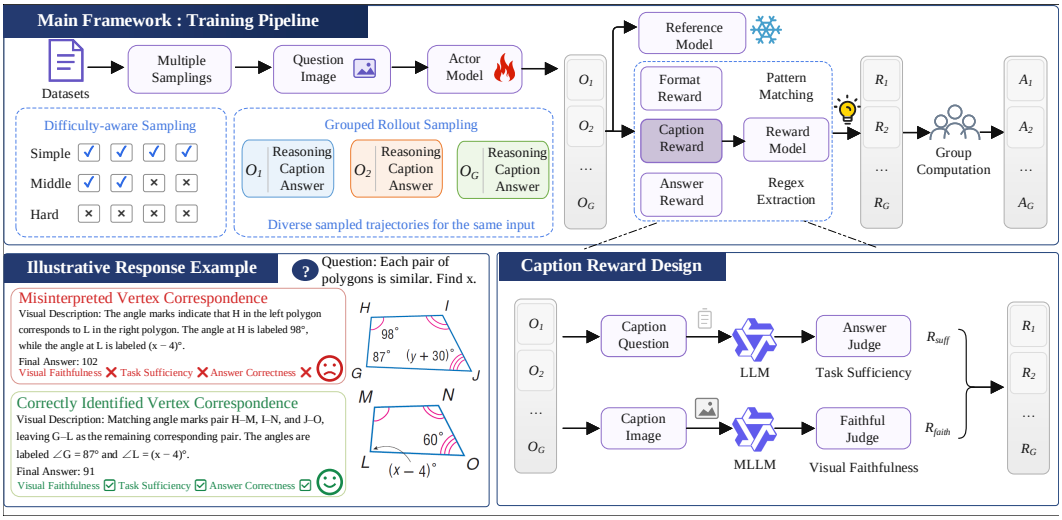


Figure 2: Overview of VERA. The top panel shows the reinforcement learning pipeline, where each response consists of reasoning, task-oriented visual evidence, and a final answer. The bottom-left panel contrasts unsupported visual descriptions with grounded, task-relevant evidence. The bottom-left panel is a schematic illustration rather than a verbatim policy rollout. It presents two possible responses to the same image–question input and displays the evidence description separately from the final answer.

3 METHOD

3.1 PRELIMINARIES

We consider a multimodal reasoning sample

$$x = (I, Q, Y), \tag{1}$$

where I, Q, and Y denote the input image, question, and ground-truth answer, respectively. The policy π_θ receives only I and Q, while Y is reserved for reward computation. For each image–question pair, the policy samples a group of responses:

$$\mathcal{O} = \{O_i\}_{i=1}^{G}, \qquad O_i \sim \pi_\theta(\cdot \mid I, Q), \tag{2}$$

where G is the group size and i indexes a sampled response. Each response is structured as

$$O_i = (Z_i, C_i, A_i), \tag{3}$$

where Z_i, C_i, and A_i denote the reasoning process, task-oriented visual evidence, and final answer, respectively. Unlike a general image caption, C_i is conditioned on the question and records only task-relevant visual information, such as objects, numerical values, attributes, and spatial relationships. It describes visual observations rather than derived solutions or an explicit final answer, allowing the evidence to be evaluated separately from the prediction. In our implementation, the policy generates reasoning, the visual-evidence description, and the final answer in this order. The description serves as an explicit target for auxiliary supervision and is not enforced as an information bottleneck for reasoning.

We characterize reliable visual evidence through visual faithfulness and task sufficiency. Following visual entailment Xie et al. (2019), let $\mathrm{Claims}(C_i)$ denote the visual propositions expressed in C_i, and let $\mathrm{Vis}(I)$ denote the visual facts supported by the image. Visual faithfulness requires

$$\mathrm{Claims}(C_i) \subseteq \mathrm{Vis}(I), \tag{4}$$

so that every visual claim in C_i is supported by the image. Task sufficiency concerns whether the evidence contains enough information to determine the answer. Motivated by evidence-only rationale evaluation DeYoung et al. (2020), let $\mathrm{Ans}(Q, C_i)$ denote the set of answers consistent with

the question and the information stated in the evidence. For a task with a single target answer, the ideal sufficiency condition is

$$\mathrm{Ans}(Q, C_i) = \{Y\}, \tag{5}$$

where semantically equivalent answer forms are treated as identical. The two conditions are complementary: visually grounded evidence may omit information required for answering, while answer-sufficient evidence may contain unsupported visual claims. VERA therefore supervises the two properties separately. Because evidence sufficiency does not ensure that the policy itself produces the correct answer Joren et al. (2025), final-answer correctness and format compliance are also evaluated.

3.2 Visual Evidence Reinforcement

We operationalize Equations 4 and 5 with two judges. The visual judge evaluates evidence against the image, while the task judge attempts to recover the answer from the question and evidence without access to the image. Both judges are instructed to reject answer-style conclusions and derived solutions presented in place of visual observations. Detailed prompts and failure-handling rules are provided in the appendix.

Visual faithfulness. The visual judge $\mathcal{J}_{\mathrm{V}}$ receives I and C_i. The corresponding reward is

$$R^{(i)}_{\mathrm{faith}} = \mathbb{I}\left[\mathcal{J}_{\mathrm{V}}(I, C_i) = 1\right], \tag{6}$$

where $\mathbb{I}[\cdot]$ is the indicator function. A positive reward indicates that the judge accepts the evidence as supported by the image.

Task sufficiency. The task judge $\mathcal{J}_{\mathrm{T}}$ receives only Q and C_i. Its predicted answer is compared with Y outside the judge:

$$R^{(i)}_{\mathrm{suff}} = \mathbb{I}\left[\mathrm{Match}\left(\mathcal{J}_{\mathrm{T}}(Q, C_i), Y\right) = 1\right], \tag{7}$$

where $\mathrm{Match}(\cdot,\cdot)$ returns 1 when the predicted and reference answers match and 0 otherwise. We use answer recovery from the question and evidence description by an image-blind judge as a model-based proxy for task sufficiency.

Answer correctness. We separately evaluate the policy's final answer:

$$R^{(i)}_{\mathrm{ans}} = \mathbb{I}\left[\mathrm{Match}(A_i, Y) = 1\right]. \tag{8}$$

Format compliance. We encourage the required response structure with

$$R^{(i)}_{\mathrm{fmt}} = \mathbb{I}\left[O_i \in \mathcal{F}\right], \tag{9}$$

where $\mathcal{F}$ denotes the valid format containing the reasoning, visual evidence, and final answer fields.

Reward aggregation. Answer correctness represents task success, while visual faithfulness, task sufficiency, and format compliance provide auxiliary supervision. The total reward is

$$R_i = \lambda_{\mathrm{ans}} R^{(i)}_{\mathrm{ans}} + \lambda_{\mathrm{faith}} R^{(i)}_{\mathrm{faith}} + \lambda_{\mathrm{suff}} R^{(i)}_{\mathrm{suff}} + \lambda_{\mathrm{fmt}} R^{(i)}_{\mathrm{fmt}}, \tag{10}$$

where the λ terms are fixed reward coefficients whose values are reported in the appendix. Following common practice in multimodal reinforcement learning with verifiable rewards (Wang et al., 2025e; Dong et al., 2025), we use a fixed scalarization of the individual reward signals throughout training. The same coefficients are used across all experiments, so that the analysis isolates the effects of the reward design itself, particularly the faithfulness and sufficiency signals, rather than differences arising from per-setting reward-weight tuning.

Policy optimization. For responses sampled from the same input, we compute group-relative advantages by normalizing rewards within the same response group. We optimize the policy using the token-level clipped objective of DAPO Yu et al. (2025). During each update, response groups are generated by a fixed policy snapshot $\pi_{\theta_{\mathrm{old}}}$. For response $O_i = (o_{i,1}, \ldots, o_{i,T_i})$, where $o_{i,<t}$ denotes the preceding tokens, the group-relative advantage and token-level importance ratio are defined as:

$$\widehat{A}_i = \frac{R_i - \mu_R}{\sigma_R + \varepsilon}, \quad \rho_{i,t}(\theta) = \frac{\pi_\theta(o_{i,t} \mid I, Q, o_{i,<t})}{\pi_{\theta_{\text{old}}}(o_{i,t} \mid I, Q, o_{i,<t})}. \tag{11}$$

where μ_R and σ_R denote the mean and standard deviation of rewards within the same response group, respectively, and $\varepsilon > 0$ is a small constant for numerical stability. The response-level advantage $\widehat{A}_i$ is shared by all tokens in O_i, and the importance ratio measures the probability change of each generated token between the updated policy and the behavior policy.

We maximize the following objective:

$$\mathcal{J}_{\text{VERA}}(\theta) = \mathbb{E}\left[\frac{1}{\sum_{i=1}^{G} T_i} \sum_{i=1}^{G} \sum_{t=1}^{T_i} \min\left(\rho_{i,t}(\theta)\widehat{A}_i, \text{clip}\left(\rho_{i,t}(\theta), 1-\epsilon_{\text{low}}, 1+\epsilon_{\text{high}}\right)\widehat{A}_i\right)\right]. \tag{12}$$

Here, $\text{clip}(u, a, b)$ restricts u to $[a, b]$, and ϵ_{low} and ϵ_{high} are the clipping margins. The expectation is taken over training samples x and response groups generated by $\pi_{\theta_{\text{old}}}$. The response-level advantage $\widehat{A}_i$ is shared by all tokens in O_i. Equation equation 12 presents the clipped policy term. In implementation, we additionally apply KL regularization with coefficient $\beta = 0.04$.

4 Experiments

We organize the experiments around three research questions:

RQ1: Does VERA improve multimodal reasoning accuracy across different multimodal reasoning model families and visual reasoning benchmarks?

RQ2: How do visual-faithfulness and task-sufficiency rewards affect training dynamics and group-relative credit assignment? **RQ3:** How do VERA's key design components contribute to performance, and how do the two evidence judges respond to automatically generated descriptions targeting different combinations of visual faithfulness and task sufficiency?

4.1 Experimental Setup

Datasets. VERA is trained on the Vision-R1-RL dataset Huang et al. (2025b). Evaluation covers eight multimodal reasoning benchmarks: MMStar Chen et al. (2024), MMMath Sun et al. (2024), WeMath Qiao et al. (2025), ChartQA Masry et al. (2022), LogicVista Xiao et al. (2024), MathVerse Zhang et al. (2024), MathVision Wang et al. (2024), and MathVista Lu et al. (2024). These benchmarks cover general multimodal understanding, chart reasoning, visual logic, and mathematical reasoning over structured visual inputs. Before reinforcement learning, we apply a fixed difficulty-aware preprocessing procedure to the Vision-R1-RL training set. The resulting training pool $\mathcal{D}_{\text{train}}$ contains examples whose reference rollouts include both correct and incorrect responses, providing relative optimization signals for group-based policy optimization.

Models and baselines. Experiments cover five multimodal reasoning backbones with different architectures and scales: Vision-R1-7B, Qwen3-VL-2B, Qwen3-VL-4B, Qwen3-VL-8B, and InternVL3.5-8B Wang et al. (2025d). We compare VERA with representative multimodal reasoning and reinforcement learning approaches, including Qwen2.5-VL Bai et al. (2025), Visionary-R1 Xia et al. (2025), AdaTooler-V Wang et al. (2025a), SophiaVL-R1 Fan et al. (2025a), VL-Rethinker Wang et al. (2025b), VL-RLCS, Wei et al. (2025), ReVisual-R1 Chen et al. (2025a), Vision-R1 Huang et al. (2025b), and STEP3-VL-10B Huang et al. (2026). All baselines are evaluated with the same prompt templates, decoding configurations, maximum generation length, and answer extraction rules.

Evaluation metrics. Each evaluation example is decoded independently three times under the same sampling configuration. Avg@3 is computed by averaging the three binary correctness values for each example and then averaging across the benchmark. Generation efficiency is measured by the average number of generated output tokens from the same responses.

Model	Size	MMStar		ChartQA		WeMath		MMMath		LogicVista		MathVerse		MathVision		MathVista	
		ACC	Token	ACC	Token	ACC	Token	ACC	Token	ACC	Token	ACC	Token	ACC	Token	ACC	Token
Methods based on Qwen2.5-VL series																	
Visionary-R1	3B	35.4	174	68.1	120	10.6	456	4.9	728	19.8	215	10.0	494	24.6	419	52.4	202
Qwen2.5-VL	7B	60.3	138	85.6	90	31.7	341	47.8	573	40.4	278	45.7	363	15.4	390	75.6	192
AdaTooler-V	7B	62.3	181	86.8	134	64.1	377	49.2	588	40.8	313	58.5	386	16.1	422	73.0	226
SophiaVL-R1	7B	59.0	166	85.6	83	33.5	365	46.6	592	36.6	292	46.2	348	17.1	422	73.5	188
VL-Rethinker	7B	65.8	164	88.8	114	68.9	334	60.8	642	44.6	267	67.1	351	22.3	425	80.5	195
VL-RLCS	7B	36.1	369	87.9	292	8.4	587	57.8	1006	25.4	508	28.8	647	14.1	664	80.9	417
ReVisual-R1	7B	63.6	1588	87.0	707	65.9	2788	59.5	564	51.5	6693	68.2	3387	36.8	7121	80.5	2020
Vision-R1	7B	64.1	1952	82.2	604	77.1	5755	52.5	7191	46.2	5959	56.8	3474	22.6	4709	73.7	1642
+ VERA	7B	64.3	1799	85.6	463	81.7	5456	55.2	6807	50.8	5510	58.5	3002	25.3	4073	74.4	1101
Multimodal reasoning large language models																	
Qwen3-VL-2B	2B	63.4	3251	72.2	1552	70.1	6971	79.4	9288	52.2	18093	62.1	7510	24.3	15315	70.5	4333
+ VERA	2B	61.5	869	86.1	900	77.2	4662	80.0	8618	53.0	12911	70.0	5845	27.9	12101	82.2	2819
Qwen3-VL-4B	4B	71.0	1499	85.1	1491	82.9	4210	84.0	4196	63.8	13625	74.5	5635	31.9	12664	80.6	3751
+ VERA	4B	75.6	1005	88.0	561	85.9	1806	85.0	3529	69.1	4531	80.6	2079	39.8	9630	81.1	1328
Qwen3-VL-8B	8B	71.4	1950	85.4	911	83.2	3822	85.7	7609	64.2	11002	76.8	4533	32.8	11052	84.3	2884
+ VERA	8B	71.9	1015	88.2	926	83.3	2859	87.2	5068	67.2	8165	77.9	4074	34.1	8673	87.8	1997
InternVL3.5	8B	70.7	2148	90.0	739	78.4	3897	77.5	7287	57.3	8231	71.2	4725	29.6	12004	83.6	2554
+ VERA	8B	80.6	1467	92.2	735	88.5	3592	87.8	5598	80.4	4355	86.4	4595	49.0	6749	84.9	2444
STEP3-VL	10B	73.9	1463	90.4	463	64.4	3678	90.7	7076	64.5	4579	71.3	4856	38.4	10892	96.7	2592

Table 1: Performance across eight multimodal reasoning benchmarks. ACC denotes Avg@3 answer accuracy, computed as the mean correctness of three independently sampled responses per example, and Token denotes the average number of generated output tokens across those responses.

Implementation details. DAPO is used as the underlying GRPO-based policy optimization algorithm. During training, the policy samples eight responses per prompt with temperature 1.0 and top-$p = 1.0$. We use a learning rate of 1×10^{-6}, a maximum completion length of $32{,}768$, and a per-device batch size of 1. Training is implemented with ms-swift Zhao et al. (2025) and DeepSpeed ZeRO-3 on eight NVIDIA A800 GPUs. Qwen3.5-397B-A17B Qwen Team (2026) is used as the external judge, with separate prompts for visual faithfulness and task sufficiency evaluation. The complete judge prompts, parsing rules, and failure-handling procedures are included in the appendix.

4.2 Overall Performance (RQ1)

Table 1 evaluates VERA on five multimodal reasoning backbones and eight benchmarks spanning general multimodal understanding and visual mathematical reasoning. Compared with the corresponding baselines, VERA improves average accuracy across all evaluated model families, scales, and benchmark categories. For example, the average accuracy of Qwen3-VL-4B increases from 71.7% to 75.6%, while that of InternVL3.5-8B increases from 69.8% to 81.2%. These consistent gains indicate that the benefit of evidence-level supervision is not limited to a particular backbone architecture or type of visual reasoning task.

These gains are not accompanied by longer outputs. Across the evaluated backbones, the complete VERA training configuration reduces output-token consumption by 9.8%–48.0% while improving average accuracy. The token reductions characterize the full optimization setup and should not be attributed solely to the visual-faithfulness and task-sufficiency rewards. Overall, these results show that VERA improves average multimodal reasoning accuracy across the evaluated model families and benchmarks while producing shorter outputs.

4.3 Effects of Evidence-Level Rewards (RQ2)

Training dynamics. Figure 3 shows the evolution of the answer, task-sufficiency, visual-faithfulness, and format rewards during reinforcement learning. Across all three backbones, both evidence-level rewards increase during training. The visual-faithfulness reward rises earlier, whereas

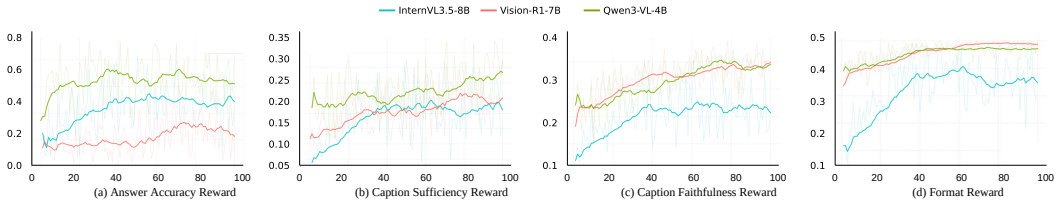


Figure 3: Training dynamics of VERA on three representative backbones. (a) Answer reward, (b) task-sufficiency reward, (c) visual-faithfulness reward, and (d) format reward. The evidence-level rewards increase alongside the answer reward, indicating that the policy obtains higher scores under the reward functions used for optimization. These curves characterize optimization behavior rather than independently validating evidence quality.

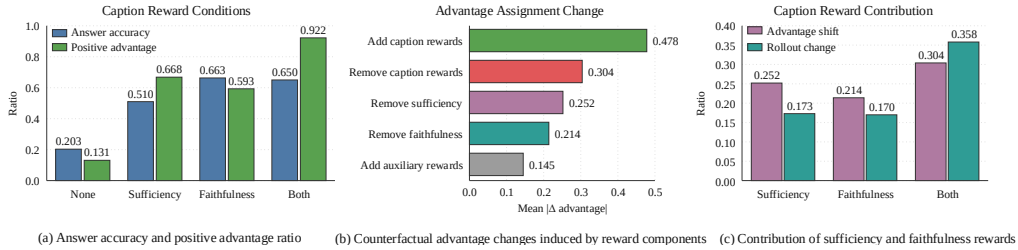


Figure 4: Effect of evidence-level rewards on group-relative credit assignment for InternVL3.5-8B. (a) Answer accuracy and positive-advantage ratios after grouping sampled responses by their realized task-sufficiency and visual-faithfulness rewards. (b) Mean absolute changes in group-normalized advantages after counterfactually adding or removing reward components while holding the sampled responses fixed. (c) Mean absolute advantage shifts and advantage-sign flip rates after removing task sufficiency, visual faithfulness, or both.

the task-sufficiency reward improves more gradually. These trends characterize how the policy responds to the combined training objective. Since the evidence curves are computed by the same judges that provide the training rewards, they do not constitute an independent human evaluation of post-training evidence quality.

Reward-induced credit assignment. Figure 4 further shows how these rewards reshape policy optimization. The positive-advantage ratio increases from 13.1% for responses satisfying neither evidence criterion to 66.8% with task sufficiency, 59.3% with visual faithfulness, and 92.2% when both criteria are satisfied. This difference is not explained by answer accuracy alone: faithfulness-only responses achieve 66.3% accuracy but receive positive advantages in only 59.3% of cases, whereas responses satisfying both criteria achieve 65.0% accuracy and a 92.2% positive-advantage ratio. Counterfactual recomputation confirms that the evidence rewards directly alter credit assignment while keeping the sampled trajectories fixed. Removing task sufficiency or visual faithfulness produces mean absolute advantage changes of 0.252 and 0.214, respectively, while removing both increases the change to 0.304 and flips the advantage sign for 35.8% of trajectories. These results show that the evidence rewards change which sampled trajectories receive positive group-relative advantages beyond answer correctness alone.

4.4 Design Ablations and Controlled Judge Diagnostics (RQ3)

RQ3 examines the contribution of VERA's key design components and characterizes the behavior of the two evidence judges on an automatically generated controlled diagnostic set.

Component ablations. Table 2 evaluates the contribution of the reward components. Removing the evidence-level reward reduces performance even when the overall response structure is retained, showing that intermediate evidence supervision provides information beyond final-answer correctness. Removing either task sufficiency or visual faithfulness also reduces performance, indicating

Ablation Variant	MMStar		ChartQA		WeMath		MMMath		LogicVista		MathVerse		MathVision		MathVista	
	ACC	Tokens	ACC	Tokens	ACC	Tokens	ACC	Tokens	ACC	Tokens	ACC	Tokens	ACC	Tokens	ACC	Tokens
w/o R_{ans}	62.1	3471	78.8	1247	73.9	4026	74.6	6743	55.7	8812	68.2	4291	31.8	5738	72.9	3215
w/o R_{fmt}	72.9	2113	84.9	978	82.0	3208	81.1	5816	64.5	7491	76.8	3627	35.7	9231	76.2	3479
w/o Evidence	70.8	1698	83.1	886	79.4	2879	78.8	5312	59.6	7158	72.7	3389	36.8	8150	78.2	2287
w/o R_{suff}	67.8	1423	82.5	712	78.1	2314	79.5	4516	60.4	6078	73.2	2854	31.2	8151	79.6	1417
w/o R_{faith}	68.5	1626	83.7	859	79.8	2816	80.8	5217	61.5	6926	74.4	3261	32.0	9469	80.4	2434
VERA	75.6	1005	88.0	561	85.9	1806	85.0	3529	69.1	4531	80.6	2079	39.8	9630	81.1	1328

Table 2: Ablation study of VERA on eight multimodal reasoning benchmarks using Qwen3-VL-4B. The w/o Evidence Rewards variant jointly removes R_{faith} and R_{suff} while retaining the structured response format.

Evidence Condition	Target R_{suff}	Target R_{faith}	N	Sufficiency Target Agreement (%)	Faithfulness Target Agreement (%)
Faithful & Sufficient	1	1	800	89.8	91.2
Faithful but Insufficient	0	1	800	84.0	90.1
Unfaithful but Sufficient	1	0	800	88.6	85.0
Unfaithful & Insufficient	0	0	800	86.2	87.4

Table 3: Target-label agreement of the task-sufficiency and visual-faithfulness judges on automatically generated controlled descriptions.

that the two dimensions provide complementary supervision. Visual faithfulness encourages evidence grounded in the image, whereas task sufficiency encourages the preservation of information needed to solve the target problem.

Controlled judge diagnostics. We construct a controlled synthetic set to examine whether the two judge roles respond to their intended evidence dimensions. We randomly sample 800 image–question–answer instances from the ChartQAPro test split Masry et al. (2025). For each instance, GPT-5.6 Sol OpenAI (2026) generates four task-oriented visual descriptions corresponding to the four combinations of visual faithfulness and task sufficiency, yielding 3,200 descriptions in total. The condition specified in each generation prompt serves as the intended label, and the generated descriptions are not independently verified by human annotators.

We evaluate the locally deployed Qwen3.5-397B-A17B using the same judge interfaces and parsing procedures employed during VERA training. The task-sufficiency judge receives the question and description without access to the image, and its recovered answer is compared with the reference answer. The visual-faithfulness judge receives the image and description and returns a binary grounding decision. As shown in Table 3, agreement with the intended conditions ranges from 84.0% to 89.8% for task sufficiency and from 85.0% to 91.2% for visual faithfulness. These results characterize judge behavior on controlled synthetic descriptions. Because the intended labels are not independently verified and the descriptions are not sampled from the final VERA policies, the evaluation may conflate judge errors with generation errors and does not rule out adaptation to judge-specific biases or shortcuts.

5 Conclusion

In this work, we show that final-answer supervision alone is insufficient for reliable multimodal reasoning, as it cannot distinguish failures in visual evidence acquisition from subsequent reasoning errors. We propose VERA, a visual evidence reinforcement framework that explicitly supervises task-relevant visual evidence through two complementary dimensions: visual faithfulness and task sufficiency. Across the five evaluated backbones, VERA improves the eight-benchmark average accuracy while reducing benchmark-averaged output-token consumption by 9.8%–48.0%. Counterfactual analyses characterize how the evidence rewards reshape group-relative credit assignment, while controlled diagnostics examine the two judge roles across predefined combinations of visual faithfulness and task sufficiency.

ETHICS STATEMENT

This study uses publicly available multimodal reasoning benchmarks and visual question-answering datasets. It does not involve human-subject experiments, clinical deployment, or the collection of personally identifiable information.

The proposed VERA framework relies on pretrained multimodal models and model-based evaluators, which may inherit biases, hallucinations, and distributional limitations from their training data and model parameters. Although VERA encourages models to provide explicit visual evidence, the generated evidence descriptions and their automated evaluations may still contain inaccurate interpretations or incomplete explanations. In particular, errors in visual-faithfulness or task-sufficiency assessment may affect the learned reward signals and subsequent policy optimization.

We encourage responsible use of multimodal reasoning systems through transparent reporting of model configurations, evaluation procedures, and known limitations. Applications involving high-impact decisions should incorporate appropriate human oversight and domain-specific validation. Any release of code, processed datasets, and evaluation resources will follow the licenses and usage restrictions of the corresponding source datasets and pretrained models.

REPRODUCIBILITY STATEMENT

We provide the main components required to reproduce VERA, including the training data construction procedure, evidence supervision process, reward design, reinforcement-learning configuration, evaluation prompts, and experimental protocols.

The supplementary material describes the automatic construction of the controlled judge diagnostic set, including the ChartQAPro sampling procedure, GPT-5.6 Sol generation protocol, intended evidence-condition definitions, and Qwen3.5-397B-A17B evaluation procedure. Implementation details are also reported for reinforcement-learning optimization, including model architectures, hyperparameters, distributed training settings, and inference configurations.

The prompts and evaluation criteria used by the visual-faithfulness and task-sufficiency judges are provided to facilitate independent analysis of the evidence-level reward mechanism. Anonymous source code and supplementary resources will be released to support reproduction, verification, and future research on reliable multimodal reasoning.

AI USE STATEMENT

Generative AI assistants were used during the research and manuscript preparation process for purposes such as language refinement, organization, and preliminary writing suggestions. All AI-generated suggestions were treated as auxiliary input only. The authors independently verified all experimental designs, implementations, numerical results, citations, figures, and scientific conclusions.

In addition, large multimodal models were used as explicit components of the VERA framework. Specifically, multimodal models were employed for evidence evaluation, including visual-faithfulness and task-sufficiency assessment, as well as other model-assisted analysis procedures described in the methodology and experimental sections. These models were treated as algorithmic components of the proposed approach rather than sources of scientific claims. Their architectures, prompts, configurations, and evaluation protocols are documented in the corresponding sections of the paper.

GPT-5.6 Sol was accessed through the OpenAI API to generate the automatically constructed descriptions used in the controlled judge diagnostic set; these descriptions were not independently verified by human annotators. Qwen3.5-397B-A17B was locally deployed for evidence and outcome evaluation.

The authors take full responsibility for the research design, experimental validation, analysis, and the accuracy, originality, and final content of this manuscript.

REFERENCES

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

Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Jiaqi Wang, Yu Qiao, Dahua Lin, and Feng Zhao. Are we on the right way for evaluating large vision-language models? In *Advances in Neural Information Processing Systems*, volume 37, 2024. doi: 10.52202/079017-0850. URL https://arxiv.org/abs/2403.20330.

Shuang Chen, Yue Guo, Zhaochen Su, Yafu Li, Yulun Wu, Jiacheng Chen, Jiayu Chen, Weijie Wang, Xiaoye Qu, and Yu Cheng. Advancing multimodal reasoning: From optimized cold start to staged reinforcement learning, 2025a. URL https://arxiv.org/abs/2506.04207.

Yang Chen, Yufan Shen, Wenxuan Huang, Sheng Zhou, Qunshu Lin, Xinyu Cai, Zhi Yu, Jiajun Bu, Botian Shi, and Yu Qiao. Learning only with images: Visual reinforcement learning with reasoning, rendering, and visual feedback, 2025b. URL https://arxiv.org/abs/2507.20766.

Xu Chu, Xinrong Chen, Guanyu Wang, Zhijie Tan, Kui Huang, Wenyu Lv, Tong Mo, and Weiping Li. Qwen look again: Guiding vision-language reasoning models to re-attention visual information, 2025. URL https://arxiv.org/abs/2505.23558.

Jiwan Chung, Junhyeok Kim, Siyeol Kim, Jaeyoung Lee, Min Soo Kim, and Youngjae Yu. v1: Learning to point visual tokens for multimodal grounded reasoning, 2025. URL https://arxiv.org/abs/2505.18842.

DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning, 2025. URL https://arxiv.org/abs/2501.12948.

Jay DeYoung, Sarthak Jain, Nazneen Fatema Rajani, Eric Lehman, Caiming Xiong, Richard Socher, and Byron C. Wallace. ERASER: A benchmark to evaluate rationalized NLP models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 4443–4458. Association for Computational Linguistics, 2020. doi: 10.18653/v1/2020.acl-main.408. URL https://aclanthology.org/2020.acl-main.408/.

Bowen Dong, Minheng Ni, Zitong Huang, Guanglei Yang, Wangmeng Zuo, and Lei Zhang. MIRAGE: Assessing hallucination in multimodal reasoning chains of MLLM. In *Advances in Neural Information Processing Systems*, volume 38, 2025. URL https://arxiv.org/abs/2505.24238.

Kaixuan Fan, Kaituo Feng, Haoming Lyu, Dongzhan Zhou, and Xiangyu Yue. Sophiavl-r1: Reinforcing mllms reasoning with thinking reward, 2025a. URL https://arxiv.org/abs/2505.17018.

Yue Fan, Xuehai He, Diji Yang, Kaizhi Zheng, Ching-Chen Kuo, Yuting Zheng, Sravana Jyothi Narayanaraju, Xinze Guan, and Xin Eric Wang. Grit: Teaching mllms to think with images. In *Advances in Neural Information Processing Systems*, 2025b. URL https://arxiv.org/abs/2505.15879.

Chang Gao, Chujie Zheng, Xiong-Hui Chen, Kai Dang, Shixuan Liu, Bowen Yu, An Yang, Shuai Bai, Jingren Zhou, and Junyang Lin. Soft adaptive policy optimization, 2025. URL https://arxiv.org/abs/2511.20347.

Haowen Gao, Zhenyu Zhang, Liang Pang, Fangda Guo, Hongjian Dou, Guannan Lv, Shaoguo Liu, Tingting Gao, Huawei Shen, and Xueqi Cheng. Diva-grpo: Enhancing multimodal reasoning through difficulty-adaptive variant advantage, 2026. URL https://arxiv.org/abs/2603.01106.

Ailin Huang, Chengyuan Yao, Chunrui Han, Fanqi Wan, Hangyu Guo, Haoran Lv, Hongyu Zhou, Jia Wang, Jian Zhou, Jianjian Sun, Jingcheng Hu, Kangheng Lin, Liang Zhao, Mitt Huang, Song Yuan, Wenwen Qu, Xiangfeng Wang, Yanlin Lai, Yingxiu Zhao, Yinmin Zhang, Yukang Shi, Yuyang Chen, Zejia Weng, Ziyang Meng, Ang Li, Aobo Kong, Bo Dong, Changyi Wan, David Wang, Di Qi, Dingming Li, En Yu, Guopeng Li, Haiquan Yin, Han Zhou, Hanshan Zhang, Haolong Yan, Hebin Zhou, Hongbo Peng, Jiaran Zhang, Jiashu Lv, Jiayi Fu, Jie Cheng, Jie Zhou, Jisheng Yin, Jingjing Xie, Jingwei Wu, Jun Zhang, Junfeng Liu, Kaijun Tan, Kaiwen Yan, Liangyu Chen, Lina Chen, Mingliang Li, Qian Zhao, Quan Sun, Shaoliang Pang, Shengjie Fan, Shijie Shang, Siyuan Zhang, Tianhao You, Wei Ji, Wuxun Xie, Xiaobo Yang, Xiaojie Hou, Xiaoran Jiao, Xiaoxiao Ren, Xiangwen Kong, Xin Huang, Xin Wu, Xing Chen, Xinran Wang, Xuelin Zhang, Yana Wei, Yang Li, Yanming Xu, Yeqing Shen, Yuang Peng, Yue Peng, Yu Zhou, Yusheng Li, Yuxiang Yang, Yuyang Zhang, Zhe Xie, Zhewei Huang, Zhenyi Lu, Zhimin Fan, Zihui Cheng, Daxin Jiang, Qi Han, Xiangyu Zhang, Yibo Zhu, and Zheng Ge. Step3-vl-10b technical report, 2026. URL https://arxiv.org/abs/2601.09668.

Qihan Huang, Weilong Dai, Jinlong Liu, Wanggui He, Hao Jiang, Mingli Song, Jingyuan Chen, Chang Yao, and Jie Song. Boosting mllm reasoning with text-debiased hint-grpo, 2025a. URL https://arxiv.org/abs/2503.23905.

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

Hailey Joren, Jianyi Zhang, Chun-Sung Ferng, Da-Cheng Juan, Ankur Taly, and Cyrus Rashtchian. Sufficient context: A new lens on retrieval augmented generation systems. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=Jjr2Odj8DJ.

Devvrit Khatri, Lovish Madaan, Rishabh Tiwari, Rachit Bansal, Sai Surya Duvvuri, Manzil Zaheer, Inderjit S. Dhillon, David Brandfonbrener, and Rishabh Agarwal. The art of scaling reinforcement learning compute for llms, 2025. URL https://arxiv.org/abs/2510.13786.

Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. In *The Twelfth International Conference on Learning Representations (ICLR)*, 2024. URL https://openreview.net/forum?id=KUNzEQMWU7.

Mi Luo, Zihui Xue, Alex Dimakis, and Kristen Grauman. When thinking drifts: Evidential grounding for robust video reasoning, 2025a. URL https://arxiv.org/abs/2510.06077.

Ruilin Luo, Zhuofan Zheng, Yifan Wang, Xinzhe Ni, Zicheng Lin, Songtao Jiang, Yiyao Yu, Chufan Shi, Lei Wang, Ruihang Chu, Jin Zeng, and Yujiu Yang. Unlocking multimodal mathematical reasoning via process reward model, 2025b. URL https://arxiv.org/abs/2501.04686.

Ahmed Masry, Do Xuan Long, Jia Qing Tan, Shafiq Joty, and Enamul Hoque. Chartqa: A benchmark for question answering about charts with visual and logical reasoning. In *Findings of the Association for Computational Linguistics: ACL 2022*, pp. 2263–2279, 2022. doi: 10.18653/v1/2022.findings-acl.177. URL https://aclanthology.org/2022.findings-acl.177/.

Ahmed Masry, Mohammed Saidul Islam, Mahir Ahmed, Aayush Bajaj, Firoz Kabir, Aaryaman Kartha, Md Tahmid Rahman Laskar, Mizanur Rahman, Shadikur Rahman, Mehrad Shahmohammadi, Megh Thakkar, Md Rizwan Parvez, Enamul Hoque, and Shafiq Joty. ChartQAPro: A more diverse and challenging benchmark for chart question answering. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 19123–19151, Vienna, Austria, July 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.findings-acl.978. URL https://aclanthology.org/2025.findings-acl.978/.

Minheng Ni, Zhengyuan Yang, Linjie Li, Chung-Ching Lin, Kevin Lin, Wangmeng Zuo, and Lijuan Wang. Point-RFT: Improving multimodal reasoning with visually grounded reinforcement finetuning, 2025. URL https://arxiv.org/abs/2505.19702.

OpenAI. GPT-5.6: Frontier intelligence that scales with your ambition. OpenAI, July 2026. URL https://openai.com/index/gpt-5-6/. Accessed: 2026-09-26.

Miao Pan, Wangjie Gan, Jintao Chen, Wenqi Zhang, Bing Sun, Jianwei Yin, and Xuhong Zhang. Ground what you see: Hallucination-resistant mllms via caption feedback, diversity-aware sampling, and conflict regularization, 2026. URL https://arxiv.org/abs/2601.06224.

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

Qwen Team. Qwen3.5: Towards native multimodal agents, February 2026. URL https://qwen.ai/blog?id=qwen3.5.

Omar Sharif, Eftekhar Hossain, Nikhil Singh, and Patrick Ng. Disentangling perception and reasoning in multimodal llms via reward design, 2026. URL https://arxiv.org/abs/2601.00215.

Kai Sun, Yushi Bai, Ji Qi, Lei Hou, and Juanzi Li. MM-MATH: Advancing multimodal math evaluation with process evaluation and fine-grained classification. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 1358–1375, 2024. doi: 10.18653/v1/2024.findings-emnlp.73. URL https://aclanthology.org/2024.findings-emnlp.73/.

Zhongwei Wan, Zhihao Dou, Che Liu, Yu Zhang, Dongfei Cui, Qinjian Zhao, Hui Shen, Jing Xiong, Yi Xin, Yifan Jiang, Chaofan Tao, Yangfan He, Mi Zhang, and Shen Yan. Srpo: Enhancing multimodal llm reasoning via reflection-aware reinforcement learning. In *Advances in Neural Information Processing Systems*, 2025. URL https://arxiv.org/abs/2506.01713.

Chaoyang Wang, Kaituo Feng, Dongyang Chen, Zhongyu Wang, Zhixun Li, Sicheng Gao, Meng Meng, Xu Zhou, Manyuan Zhang, Yuzhang Shang, and Xiangyu Yue. Adatooler-v: Adaptive tool-use for images and videos, 2025a. URL https://arxiv.org/abs/2512.16918.

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, 2025b. URL https://arxiv.org/abs/2504.08837.

Jiahao Wang, Weiye Xu, Aijun Yang, Wengang Zhou, Lewei Lu, Houqiang Li, Xiaohua Wang, and Jinguo Zhu. Enhancing the outcome reward-based rl training of mllms with self-consistency sampling, 2025c. URL https://arxiv.org/abs/2511.10648.

Ke Wang, Junting Pan, Weikang Shi, Zimu Lu, Houxing Ren, Aojun Zhou, Mingjie Zhan, and Hongsheng Li. Measuring multimodal mathematical reasoning with MATH-Vision dataset. In *Advances in Neural Information Processing Systems*, volume 37, 2024. doi: 10.52202/079017-3014. URL https://papers.nips.cc/paper_files/paper/2024/hash/ad0edc7d5fa1a783f063646968b7315b-Abstract-Datasets_and_Benchmarks_Track.html.

Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, et al. Internvl3.5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *arXiv preprint arXiv:2508.18265*, 2025d. URL https://arxiv.org/abs/2508.18265.

Yibin Wang, Zhimin Li, Yuhang Zang, Chunyu Wang, Qinglin Lu, Cheng Jin, and Jiaqi Wang. Unified multimodal chain-of-thought reward model through reinforcement fine-tuning. In *Advances in Neural Information Processing Systems*, volume 38, 2025e. URL https://arxiv.org/abs/2505.03318.

Lai Wei, Yuting Li, Kaipeng Zheng, Chen Wang, Yue Wang, Linghe Kong, Lichao Sun, and Weiran Huang. Advancing multimodal reasoning via reinforcement learning with cold start, 2025. URL https://arxiv.org/abs/2505.22334.

Jiaer Xia, Yuhang Zang, Peng Gao, Sharon Li, and Kaiyang Zhou. Visionary-r1: Mitigating shortcuts in visual reasoning with reinforcement learning, 2025. URL https://arxiv.org/abs/2505.14677.

Yijia Xiao, Edward Sun, Tianyu Liu, and Wei Wang. Logicvista: Multimodal llm logical reasoning benchmark in visual contexts. *arXiv preprint arXiv:2407.04973*, 2024. doi: 10.48550/arXiv.2407.04973. URL https://arxiv.org/abs/2407.04973.

Ning Xie, Farley Lai, Derek Doran, and Asim Kadav. Visual entailment: A novel task for fine-grained image understanding. *arXiv preprint arXiv:1901.06706*, 2019. URL https://arxiv.org/abs/1901.06706.

Donglai Xu, Hongzheng Yang, Yuzhi Zhao, Pingping Zhang, Jinpeng Chen, Wenao Ma, Zhijian Hou, Mengyang Wu, Xiaolei Li, Senkang Hu, Ziyi Guan, Jason Chun Lok Li, and Lai Man Po. From exploration to exploitation: A two-stage entropy rlvr approach for noise-tolerant mllm training, 2025a. URL https://arxiv.org/abs/2511.07738.

Shilin Xu, Yanwei Li, Rui Yang, Tao Zhang, Yueyi Sun, Wei Chow, Linfeng Li, Hang Song, Qi Xu, Yunhai Tong, Xiangtai Li, and Hao Fei. Mixed-r1: Unified reward perspective for reasoning capability in multimodal large language models, 2025b. URL https://arxiv.org/abs/2505.24164.

Qi Yang, Bolin Ni, Shiming Xiang, Han Hu, Houwen Peng, and Jie Jiang. R-4b: Incentivizing general-purpose auto-thinking capability in mllms via bi-mode annealing and reinforce learning, 2025. URL https://arxiv.org/abs/2508.21113.

Huanjin Yao, Qixiang Yin, Jingyi Zhang, Min Yang, Yibo Wang, Wenhao Wu, Fei Su, Li Shen, Minghui Qiu, Dacheng Tao, and Jiaxing Huang. R1-sharevl: Incentivizing reasoning capabilities of multimodal large language models via share-grpo. In *Advances in Neural Information Processing Systems*, volume 38, 2025. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/73b75e4eafbdb617822bba262cc81c76-Abstract-Conference.html.

Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Weinan Dai, Tiantian Fan, Gaohong Liu, Lingjun Liu, Xin Liu, Haibin Lin, Zhiqi Lin, Bole Ma, Guangming Sheng, Yuxuan Tong, Chi Zhang, Mofan Zhang, Wang Zhang, Hang Zhu, Jinhua Zhu, Jiaze Chen, Jiangjie Chen, Chengyi Wang, Hongli Yu, Yuxuan Song, Xiangpeng Wei, Hao Zhou, Jingjing Liu, Wei-Ying Ma, Ya-Qin Zhang, Lin Yan, Mu Qiao, Yonghui Wu, and Mingxuan Wang. Dapo: An open-source llm reinforcement learning system at scale, 2025. URL https://arxiv.org/abs/2503.14476.

Renrui Zhang, Dongzhi Jiang, Yichi Zhang, Haokun Lin, Ziyu Guo, Pengshuo Qiu, Aojun Zhou, Pan Lu, Kai-Wei Chang, Peng Gao, and Hongsheng Li. Mathverse: Does your multi-modal llm truly see the diagrams in visual math problems? In *Computer Vision – ECCV 2024*, pp. 169–186, 2024. URL https://arxiv.org/abs/2403.14624.

Yi-Fan Zhang, Xingyu Lu, Shukang Yin, Chaoyou Fu, Wei Chen, Xiao Hu, Bin Wen, Kaiyu Jiang, Changyi Liu, Tianke Zhang, Haonan Fan, Kaibing Chen, Jiankang Chen, Haojie Ding, Kaiyu Tang, Zhang Zhang, Liang Wang, Fan Yang, Tingting Gao, and Guorui Zhou. Thyme: Think beyond images, 2025. URL https://arxiv.org/abs/2508.11630.

Yuze Zhao, Jintao Huang, Jinghan Hu, Xingjun Wang, Yunlin Mao, Daoze Zhang, Zeyinzi Jiang, Zhikai Wu, Baole Ai, Ang Wang, Wenmeng Zhou, and Yingda Chen. SWIFT: A scalable lightweight infrastructure for fine-tuning. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(28):29733–29735, 2025. doi: 10.1609/aaai.v39i28.35383. URL https://ojs.aaai.org/index.php/AAAI/article/view/35383.

Chujie Zheng, Shixuan Liu, Mingze Li, Xiong-Hui Chen, Bowen Yu, Chang Gao, Kai Dang, Yuqiong Liu, Rui Men, An Yang, Jingren Zhou, and Junyang Lin. Group sequence policy optimization, 2025. URL https://arxiv.org/abs/2507.18071.

Ziwei Zheng, Michael Yang, Jack Hong, Chenxiao Zhao, Guohai Xu, Le Yang, Chao Shen, and Xing Yu. Deepeyes: Incentivizing "thinking with images" via reinforcement learning. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://arxiv.org/abs/2505.14362.

A Related Work

Research on multimodal large language models has increasingly moved beyond visual question answering and general cross-modal understanding toward multi-step reasoning over complex visual inputs. Inspired by text-only reasoning paradigms such as DeepSeek-R1 DeepSeek-AI et al. (2025), recent work applies reinforcement learning to mathematical reasoning, chart understanding, spatial inference, and visual question answering. We organize the most relevant literature into four directions: optimization foundations for multimodal reinforcement learning, visual faithfulness, adaptive and agentic reasoning, and reward- and advantage-related optimization issues.

A.1 Reinforcement Learning Foundations

One major line of work focuses on establishing a stable optimization foundation for multimodal reinforcement learning, especially through cold-start initialization, curriculum-style supervision, reward design, and stabilization strategies. Vision-R1 Huang et al. (2025b) shows that when reinforcement learning is directly applied to multimodal reasoning, stable and sophisticated reasoning patterns rarely emerge naturally. Advancing Multimodal Reasoning via Reinforcement Learning with Cold Start Wei et al. (2025) and ReVisual-R1 Chen et al. (2025a) similarly argue that structured supervised initialization is not merely helpful but often essential for subsequent reinforcement learning to converge effectively. URSA Luo et al. (2025b) further demonstrates that automatically constructed large-scale multimodal mathematical chain-of-thought data and process-supervision data provide a strong basis for unlocking multimodal reasoning capabilities.

Beyond initialization, another key issue concerns reward design and optimization stability in RLVR-based multimodal training. R1-ShareVL Yao et al. (2025) addresses reward sparsity and advantage vanishing by sharing trajectories and advantage estimation across problem variants, thereby improving both the density and stability of training signals. Mixed-R1 Xu et al. (2025b) proposes a unified reward perspective that integrates multiple-choice questions, chart understanding, grounding tasks, and open-ended generation into a single post-training framework. SophiaVL-R1 Fan et al. (2025a) extends reward modeling to the reasoning process itself by combining thinking rewards with outcome rewards, while introducing confidence weighting and annealing mechanisms to reduce reward hacking. Self-Consistency Sampling Wang et al. (2025c) and two-stage entropy optimization methods Xu et al. (2025a) further improve multimodal reinforcement learning from the perspectives of trajectory consistency, noise robustness, and more stable single-rollout optimization.

A.2 Visual Faithfulness

A second line of work focuses on visual faithfulness, namely whether multimodal reasoning chains genuinely rely on visual evidence rather than merely producing plausible textual reasoning. Several studies attempt to make visual evidence usage more explicit during reasoning. GRIT Fan et al. (2025b) improves the auditability of reasoning by allowing the model to alternately generate natural language and bounding-box coordinates, thereby explicitly specifying the image regions on which its reasoning is based. At a finer granularity, v1 Chung et al. (2025) introduces a point-and-copy mechanism that enables the model to revisit critical visual patches during reasoning instead of relying solely on language-side continuation once image encoding has been completed. From the perspective of attention degradation, Qwen Look Again Chu et al. (2025) shows that as reasoning chains become longer, attention to visual tokens is progressively diluted.

Closely related to visual faithfulness is the problem of shortcut learning and hallucination. Visionary-R1 Xia et al. (2025) finds that when training depends only on question-answer pairs and

outcome rewards, models may exploit textual shortcuts from easier examples instead of acquiring genuine visual understanding. Ground What You See Pan et al. (2026) further analyzes hallucination from the perspective of training dynamics, arguing that its causes lie not only in insufficient visual anchoring during reasoning but also in inadequate exploration diversity and inter-sample update conflicts. From Sight to Insight Sharif et al. (2026) identifies insufficient visual perception, rather than purely algorithmic weakness, as a core source of failure in visual reasoning, and promotes actual use of visual information during reinforcement learning through multiple reward functions.

A.3 Adaptive and Agentic Reasoning

A third line of work extends multimodal reasoning beyond static chain-of-thought generation toward more adaptive and agentic forms of reasoning. One subdirection emphasizes reflection and controllable reasoning behavior. VL-Rethinker Wang et al. (2025b) transforms reflection from a naturally emerging phenomenon into an explicitly trainable behavior. SRPO Wan et al. (2025) further proposes reflection-aware reinforcement learning, where reflection-oriented data construction and reflection reward design encourage reflections that are not only present but also concise and targeted. R-4B Yang et al. (2025) goes one step further by treating the decision of whether to enter a thinking mode itself as a learnable policy, thereby seeking a dynamic balance between reasoning cost and performance.

Another emerging subdirection investigates tool-augmented and executable multimodal reasoning. DeepEyes Zheng et al. (2026) attempts to elicit interleaved multimodal reasoning capabilities through end-to-end reinforcement learning without relying on cold-start supervised fine-tuning, while treating grounding ability itself as an internal tool. Thyme Zhang et al. (2025) moves beyond simply re-inspecting the original image by allowing the model to autonomously generate and execute code for operations such as cropping, rotation, and enhancement, and then use the processed results to support subsequent reasoning. The RRVF framework in Learning Only with Images Chen et al. (2025b) establishes a closed loop of reasoning, rendering, and visual feedback, in which intermediate outputs are generated directly from raw images and visual consistency, rather than language annotation, serves as the training signal. AdaTooler-V Wang et al. (2025a) further elevates the decision of whether tools are needed into a learnable policy, aiming to avoid ineffective or excessive tool usage. These studies suggest that multimodal reasoning is gradually evolving from chain-like reasoning over static inputs toward an executable, verifiable, and interactive agentic paradigm.

A.4 Reward and Advantage Issues

In multimodal GRPO-style reinforcement learning, optimization failure is more commonly associated with sparse rewards, vanishing group-relative advantages, and gradient stagnation than with conventional layer-wise gradient vanishing. Sparse rewards arise when only a small fraction of sampled trajectories receive positive feedback, making exploration inefficient. Advantage vanishing occurs when rollouts within the same prompt group receive identical or nearly identical rewards, leaving little or no relative signal for policy updates.

VL-Rethinker Wang et al. (2025b) notes that identical rewards within a rollout group collapse the relative advantage to zero. R1-ShareVL Yao et al. (2025) addresses the joint problem of reward sparsity and advantage vanishing by sharing trajectories and estimating advantages across related problem variants. ReVisual-R1 Chen et al. (2025a) characterizes the resulting optimization failure as gradient stagnation and introduces PAD to filter zero-advantage samples and reweight informative trajectories. Hint-GRPO Huang et al. (2025a) improves data utilization when all responses in a group are incorrect through adaptive hints, whereas DIVA-GRPO Gao et al. (2026) uses difficulty-adaptive variants to preserve within-group reward diversity for both overly easy and overly difficult problems. SophiaVL-R1 Fan et al. (2025a) further stabilizes training through process-aware reward shaping and confidence-weighted supervision. Collectively, these studies indicate that effective multimodal reinforcement learning depends not only on reward magnitude, but also on maintaining informative reward variation within each optimization group.

Benchmark	Size	Format	Primary capability
MMStar	1500	Mixed	General multimodal understanding across six capability domains and fine-grained dimensions.
MMMath	5929	Open-ended	Multimodal mathematical reasoning with outcome- and process-level annotations.
WeMath	1740	Open-ended	Step-by-step visual mathematical reasoning across hierarchical knowledge concepts.
ChartQA	2500	QA	Reading and reasoning over structured charts, including automatically generated questions.
LogicVista	448	Multiple choice	Deductive, inductive, numerical, spatial, and mechanical reasoning in visual contexts.
MathVerse	3940	Open-ended	Diagram understanding under six information configurations.
MathVision	304	Open-ended	Competition-level visual mathematics across 16 disciplines and five difficulty levels.
MathVista	1000	Mixed	Broad mathematical reasoning over diverse visual contexts.

Table 4: Summary of the evaluation benchmarks. Dataset descriptions follow the corresponding benchmark papers.

B Supplementary Experimental Details

This appendix provides the implementation and evaluation details required to reproduce VERA, including benchmark statistics, training-data filtering, reward computation, optimization settings, and additional analyses. It also presents controlled caption examples that disentangle visual faithfulness from task sufficiency.

B.1 Construction of the Controlled Judge Diagnostic Set

We randomly sample 800 questions from the test split of ChartQAPro Masry et al. (2025). Each source instance consists of a chart image, a question, and one or more reference answers.

For each instance $x_k = (I_k, Q_k, Y_k)$, GPT-5.6 Sol OpenAI (2026) is prompted to construct four task-oriented visual descriptions:

$$C_k^{(1,1)}, \quad C_k^{(1,0)}, \quad C_k^{(0,1)}, \quad C_k^{(0,0)},$$

where the first indicator specifies the intended visual faithfulness and the second specifies the intended task sufficiency. This paired construction produces 3,200 descriptions, with 800 descriptions targeting each condition.

The generation instructions require faithful descriptions to contain only chart-supported visual claims and sufficient descriptions to preserve the information needed to recover the reference answer from the question and description alone. Faithful-but-insufficient descriptions omit necessary information without introducing unsupported claims. Unfaithful-but-sufficient descriptions retain the necessary information while introducing an unsupported visual claim. Unfaithful-and-insufficient descriptions combine unsupported claims with missing or incorrect task-relevant information. The generator is instructed to exclude condition labels, explicit answer conclusions, calculations, and derived solutions from all descriptions.

We evaluate Qwen3.5-397B-A17B on these descriptions using the same judge protocols as in VERA. The visual-faithfulness judge receives the image and description, whereas the task-sufficiency judge receives only the question and description, with its recovered answer compared against the reference answer outside the judge. This diagnostic evaluation examines judge behavior across the four intended evidence conditions. The descriptions are generated automatically without human review; their condition labels therefore represent generation targets rather than independently verified ground truth.

B.2 Evaluation Benchmarks

Table 4 summarizes the eight benchmarks used in the main evaluation. We use the official evaluation split for each benchmark unless stated otherwise.

B.3 Training Data Construction and Difficulty-Aware Filtering

We apply difficulty-aware preprocessing to remove rollout groups that provide no relative learning signal. Table 5 reports the complete filtering protocol and the resulting training-set statistics.

Starting from Vision-R1-rl, we remove examples with invalid answers, unreadable or corrupted images, duplicate questions, unparsable fields, or sequences exceeding the maximum length. We

Item	Value	Description
Training dataset	Osilly/Vision-R1-rl	Public multimodal mathematical reasoning dataset used as the original training corpus.
Original number of examples	11,312	Number of examples before data cleaning and difficulty-aware filtering.
Reference model	Qwen3-VL-2B-Thinking	Model used to estimate the empirical difficulty of each training example.
Rollouts per example	8	Number of independently sampled responses generated for each example during difficulty estimation.
Filtering temperature / top-p	1.0 / 0.95	Sampling configuration used to generate the responses for difficulty estimation.
Uniformly correct groups removed	2,036	Examples for which all eight sampled responses produced the correct final answer.
Uniformly incorrect groups removed	2,722	Examples for which none of the eight sampled responses produced the correct final answer.
Malformed or corrupted samples removed	147	Examples containing invalid or unreadable images, missing reference answers, duplicated records, or fields that could not be parsed.
Final number of training examples	6,407	Examples retained for policy optimization after data cleaning and difficulty-aware filtering.
Retention rate	56.64%	Ratio between the final number of retained examples and the original number of examples, i.e., $6{,}407/11{,}312$.
Shared across backbones	Yes	All evaluated backbone models use the identical filtered training pool to ensure a controlled comparison.

Table 5: Difficulty-aware data construction and filtering statistics. The filtering counts are based on the final training pool.

then discard uniformly correct and uniformly incorrect rollout groups according to the reference-model outputs. All backbone models are trained on the same filtered pool to ensure a controlled comparison.

B.4 Failure Annotation Protocol

The failure analysis in the main text distinguishes errors in visual recognition from errors in subsequent reasoning. We use the following operational definitions:

- **Visual-recognition error:** At least one task-relevant visual claim used by the response contradicts the image, introduces an unsupported visual fact, or omits visual evidence required by the attempted reasoning.
- **Reasoning error:** The task-relevant visual evidence is identified correctly, but a subsequent mathematical or logical inference is incorrect.

For the mutually exclusive visualization in Figure 1(a), any response containing a visual-recognition error is assigned to the visual-recognition category. A response is assigned to the reasoning category only when its task-relevant visual evidence is correct but a subsequent inference is incorrect.

C Prompts and Reward Implementation

C.1 Training Output Format

Figure 5 presents the instruction used during training to separate the model's reasoning process, task-relevant visual evidence, and final answer into distinct fields.

C.2 Task-Sufficiency Judge

Figure 6 shows the prompt used by the task-sufficiency judge to determine whether the generated caption, together with the problem statement, contains sufficient information to recover the reference answer without access to the image.

C.3 Visual-Faithfulness Judge

Figure 7 presents the prompt used by the visual-faithfulness judge to verify whether every substantive visual claim in the caption is directly grounded in the input image.

MRLM Instruction for Training

A conversation between User and Assistant. The user asks a question about a mathematical diagram or chart, and the Assistant solves it. The Assistant must strictly follow this output structure:
<think> Step-by-step mathematical reasoning process. </think>
<caption> Provide a structured and precise description of the image. </caption>
<answer> Final answer wrapped with \boxed{}. </answer>
The <caption> must contain only observable visual information. Describe positions, labels, quantities, and geometric relationships that are directly supported by the image. Do not include calculations, derived conclusions, intermediate reasoning, or the final answer in the caption.

Figure 5: Training instruction used to elicit reasoning, visual evidence, and the final answer in separate fields.

Instruction for the Task-Sufficiency Judge

The model is provided with two inputs:
(1) a caption containing task-relevant visual information;
(2) a mathematical problem associated with the image.
The original image is not provided. Solve the problem using only the information stated in the caption and the problem. Do not infer missing visual information or reconstruct omitted image details.
Do not evaluate whether the caption is faithful to the image. If the caption contains the final answer, explicit calculations, completed reasoning, or derived conclusions, output INVALID_CAPTION. Otherwise, solve the problem and return only the final answer in the format \boxed{}.
Caption: {caption}
Problem: {problem}

Figure 6: Prompt used to evaluate whether a caption contains enough information to solve the associated problem.

C.4 Outcome Judge

Figure 8 shows the prompt used by the outcome judge to determine whether the normalized prediction is semantically equivalent to the reference answer.

C.5 Judge Output Parsing and Failure Handling

All judge requests are processed by a locally deployed Qwen3.5-397B-A17B model running on a separate server equipped with 16 NVIDIA H800 GPUs. The judge server communicates with the training workers over the same intranet and does not rely on external APIs. This setup provides stable, high-throughput reward evaluation while keeping the entire training pipeline within the local computing environment.

Before parsing, leading and trailing whitespace and redundant line breaks are removed from the judge output. For the visual-faithfulness judge, the parser searches for a valid occurrence of \boxed{0} or \boxed{1}. Whitespace between \boxed and the braces is tolerated. Additional explanatory text does not invalidate the output when exactly one valid boxed score is present. Outputs containing no valid boxed score or multiple conflicting boxed scores are treated as invalid.

For the task-sufficiency judge, an output containing INVALID_CAPTION is directly assigned a task-sufficiency reward of zero. Otherwise, a single boxed answer is extracted and normalized using the same procedure applied to the final prediction and reference answer. The normalized answer is then passed to the outcome-matching procedure.

For the outcome judge, leading and trailing whitespace is removed, and the output is matched case-insensitively against YES and NO. Any other output is treated as invalid. Judge requests are not resampled at the reward-computation level. Therefore, the maximum number of reward-level retries is zero. Any unresolved timeout, request failure, or parsing failure receives a conservative fallback reward of zero.

Instruction for the Visual-Faithfulness Judge

You are a visual-faithfulness evaluation assistant.
You will receive:
(1) an image associated with a mathematical problem;
(2) a caption describing visual information from the image.
Determine whether every substantive visual claim in the caption is directly supported by the image. Do not judge whether the caption is sufficient to solve the problem, and do not penalize the caption for omitting visible details.
Return `\boxed{1}` if all stated visual claims are supported by the image and the caption contains only directly observable information.
Return `\boxed{0}` if the caption contains any contradiction, unsupported or invented detail, incorrect object, label, number, or visual relationship, or if it includes calculations, derived conclusions, intermediate results, or a final answer that is not directly visible in the image.
An incomplete but visually correct caption should receive `\boxed{1}`.
Caption: {`caption`}
Return only the score in the format `\boxed{}`.

Figure 7: Prompt used to evaluate whether the caption is grounded in directly observable image evidence.

Instruction for the Outcome Judge

Determine whether the prediction matches the reference answer. Superficial lexical overlap or accidental string matching should not be considered a valid match. Return **YES** if a valid match exists; otherwise, return **NO**.
Prediction: {`normalized_prediction`}
Reference Answers: {`normalized_golden`}

Figure 8: Prompt used to compute the answer-correctness reward.

This conservative fallback prevents unavailable, ambiguous, or malformed judge responses from introducing unsupported positive training signals.

C.6 Reward Composition

All reward components are binary. Answer correctness receives the largest weight because final-answer accuracy remains the primary objective, while task sufficiency, visual faithfulness, and output-format compliance provide complementary supervision for reliable and task-relevant visual evidence.

For response i, the total reward is defined as

$$R^{(i)} = R^{(i)}_{\text{ans}} + 0.5R^{(i)}_{\text{suff}} + 0.5R^{(i)}_{\text{faith}} + 0.5R^{(i)}_{\text{fmt}}. \tag{13}$$

Each component belongs to $\{0, 1\}$, and the resulting total reward lies in the interval $[0, 2.5]$.

Thus, answer correctness contributes 1.0, whereas each auxiliary term contributes 0.5. An invalid caption does not introduce an additional negative reward. Instead, an output classified as `INVALID_CAPTION` receives $R_{\text{suff}} = 0$. Captions containing leaked answers, explicit calculations, derived conclusions, or unsupported visual statements are evaluated independently by the visual-faithfulness judge and receive $R_{\text{faith}} = 0$ when such violations are detected.

The individual rewards and their weighted sum are not manually clipped, normalized, or rescaled before group-relative advantage estimation. The total rewards are then converted into group-relative advantages using the reward statistics of responses sampled for the same prompt, following the DAPO implementation used in our experiments.

Condition	Action	Implementation detail
Faithfulness output is valid	Parse as a binary score	Accept exactly one whitespace-tolerant occurrence of `\boxed{0}` or `\boxed{1}`.
Sufficiency output is a valid answer	Normalize and match	Extract one boxed answer, normalize it, and compare it with the normalized reference answer using the outcome-matching procedure.
Sufficiency output is `INVALID_CAPTION`	Assign $R_{\text{suff}} = 0$	The response remains in the sampled group, but it receives no task-sufficiency reward.
Output contains multiple values	Assign zero	Multiple or conflicting boxed values are treated as ambiguous rather than selecting the first or last occurrence.
Output contains additional prose	Apply tolerant parsing	Additional prose is ignored when exactly one valid target value is present; otherwise, the output is treated as invalid.
Judge request fails or times out	Assign zero	No reward-level retry is performed. Any unresolved timeout, request failure, or parsing failure receives a fallback reward of zero.
Malformed training response	Apply field-specific zero rewards	Missing or malformed output tags yield $R_{\text{fmt}} = 0$. A missing or unparseable `<caption>` yields $R_{\text{suff}} = R_{\text{faith}} = 0$, while a missing or unparseable `<answer>` yields $R_{\text{ans}} = 0$.

Table 6: Judge-output parsing and failure-handling rules.

Reward term	Weight	Definition and edge cases
Answer correctness R_{ans}	1.0	Equals 1 when the outcome judge determines that the normalized prediction matches a reference answer, and 0 otherwise. Invalid, missing, or unparseable answers receive 0.
Task sufficiency R_{suff}	0.5	Equals 1 when the task-sufficiency judge can recover the reference answer using only the problem and caption. An `INVALID_CAPTION`, malformed output, failed request, or incorrect recovered answer receives 0.
Visual faithfulness R_{faith}	0.5	Equals 1 only when every substantive visual claim is supported by the image and the caption contains no prohibited reasoning, derived conclusion, or leaked answer. Otherwise, it equals 0.
Format reward R_{fmt}	0.5	Equals 1 when the response contains correctly ordered and parseable `<think>`, `<caption>`, and `<answer>` fields, with the final answer enclosed in `\boxed{}`. Otherwise, it equals 0.
Invalid-caption handling	—	No separate penalty term is introduced. An invalid caption receives $R_{\text{suff}} = 0$, while its visual faithfulness is independently evaluated by the visual-faithfulness judge.
Total reward	—	$R = R_{\text{ans}} + 0.5R_{\text{suff}} + 0.5R_{\text{faith}} + 0.5R_{\text{fmt}}$, with $R \in [0, 2.5]$.

Table 7: Reward components and coefficients used by VERA.

D Training Configuration

D.1 Hardware and Software Environment

Table 8 summarizes the hardware resources and major software dependencies used for policy training, rollout generation, and reward evaluation.

D.2 Training Hyperparameters

Table 9 summarizes the policy-optimization and rollout-generation hyperparameters used in the main experiments, where eight responses are sampled for each training example.

E Qualitative Case Analysis

The following examples illustrate the four logical combinations of caption faithfulness and task sufficiency. A caption is *faithful* when every substantive claim is directly supported by the image. It is *sufficient* when, together with the problem statement, it provides enough information to recover the reference answer without access to the image. Sufficiency does not imply faithfulness: a caption may contain all information needed to solve the task while also introducing an unsupported visual claim. These categories describe logical information content; in the implemented reward, captions containing explicit calculations or derived conclusions are additionally rejected as `INVALID_CAPTION`.

Component	Value	Notes
GPU type and count	8× NVIDIA A800-SXM4-80GB	Eight GPUs, each with 80 GB of memory, were used for policy training and rollout generation.
CPU	Intel Xeon Platinum 8358P, 2.60 GHz	The training server provides 128 CPU cores.
System memory	1007 GB	Host memory available for data loading, rollout management, and distributed training.
NVIDIA driver	580.65.06	NVIDIA driver installed on the policy-training server.
CUDA driver compatibility	13.0	Maximum CUDA version supported by the installed NVIDIA driver, as reported by `nvidia-smi`.
PyTorch CUDA build	12.8	CUDA version against which the installed PyTorch package was built.
Python	3.12	Exact interpreter version used in the training environment.
PyTorch	2.10.0	Deep-learning framework used for model training and optimization.
TorchVision	0.25.0	Vision-processing components used with PyTorch.
Transformers	4.57.4	Hugging Face model, configuration, and tokenizer implementation.
Tokenizers	0.22.2	Tokenization backend used by Transformers.
Accelerate	1.13.0	Distributed training and device-management utilities.
vLLM	0.17.1	Inference engine used for rollout generation during reinforcement learning.
DeepSpeed	0.18.8	Distributed optimization framework used with ZeRO Stage 3.
Training framework	ms-swift 4.2.3	Framework used to implement multimodal GRPO/DAPO training.
TRL	0.28.0	Reinforcement-learning trainer dependency used by the training framework.
Datasets	3.6.0	Dataset loading, transformation, and preprocessing library.
FlashAttention	2.8.3	Memory-efficient attention implementation used in the training environment.
Ray	2.54.1	Distributed execution framework used to coordinate training and inference workers.
Judge serving framework	vLLM 0.17.1	Qwen3.5-397B-A17B was deployed offline on a separate server equipped with 16 NVIDIA H800 GPUs. The judge server and training workers communicated within the same intranet.

Table 8: Hardware and software environment used for policy training, rollout generation, and reward evaluation.

F LIMITATIONS

Our experiments primarily use English- and Chinese-language multimodal benchmarks; performance in other languages remains untested. The caption-supervision analysis also focuses on structured visual-reasoning inputs, particularly mathematical diagrams and charts. Its applicability to natural images, medical images, and complex real-world scenes therefore requires further evaluation.

The faithfulness and sufficiency rewards depend on model-based judges. Although the two judges target distinct properties, they may still exhibit systematic biases, correlated errors, or sensitivity to uncommon but valid evidence formulations. The controlled diagnostic set evaluates their behavior on constructed examples but does not provide independent human validation. We also do not audit descriptions generated by the final VERA policies for exploitation of the reward judges. The reported target-label agreement rates and reward trajectories do not rule out policy adaptation to judge-specific biases or shortcuts.

VERA is evaluated primarily with DAPO. Other recent policy-optimization objectives, including CISPO Khatri et al. (2025), GSPO Zheng et al. (2025), and SAPO Gao et al. (2025), may induce different training dynamics. The current system also does not use external image-processing or verification tools. Operations such as cropping, zooming, rendering, or programmatic measurement may be beneficial for small labels, dense charts, and visually ambiguous inputs. Because the evidence description is generated after the reasoning trace, the current experiments do not establish that it causally guides the preceding reasoning. The description may partly summarize evidence retro-

Parameter	Value	Description
seed / data_seed	42 / 42	Random seeds used for model training and dataset sampling.
torch_dtype	bfloat16	Numerical precision used for policy training.
num_generations (main training)	8	Number of independently sampled responses generated for each training example.
temperature	1.0	Sampling temperature used during rollout generation.
top_p	1.0	Nucleus-sampling threshold used during rollout generation.
max_completion_length	32,768	Maximum number of generated tokens for each rollout.
num_train_epochs	2	Number of passes over the filtered training dataset.
learning_rate	1×10^{-6}	Learning rate of the policy optimizer.
per_device_train_batch_size	1	Number of training examples processed on each GPU per micro-batch.
gradient_accumulation_steps	16	Number of micro-batches accumulated before each optimizer update.
generation_batch_size	64	Number of rollout responses processed in each generation phase.
steps_per_generation	16	Number of optimization steps performed for each generation phase.
loss_type	dapo	DAPO policy-optimization objective implemented in ms-swift.
epsilon / epsilon_high	0.20 / 0.28	Lower and upper clipping coefficients for the policy probability ratio.
beta	0.04	Coefficient applied to KL regularization.
weight_decay	0.10	Weight-decay coefficient used by the optimizer.
adam_beta1 / adam_beta2	0.90 / 0.95	First- and second-order moment coefficients of the Adam optimizer.
adam_epsilon	1×10^{-8}	Numerical-stability constant of the Adam optimizer.
max_grad_norm	1.0	Maximum gradient norm used for gradient clipping.
warmup_ratio	0	No learning-rate warmup is applied.
num_iterations	1	Number of policy-update iterations performed for each rollout batch.
dynamic_sample	true	Enables resampling when all responses in a rollout group receive the same answer-correctness reward.
max_resample_times	3	Maximum number of dynamic-resampling attempts for an uninformative rollout group.
overlong_filter	true	Excludes responses truncated at the maximum completion length from policy optimization.
soft_cache_length	4,096	Length threshold used by the soft overlength mechanism.
shuffle_dataset	true	Shuffles the training examples before batching.
dataloader_drop_last	true	Discards the final incomplete training batch.
vllm_tensor_parallel_size	4	Tensor-parallel degree used by vLLM during rollout generation.
vllm_gpu_memory_utilization	0.50	Maximum fraction of GPU memory allocated to the vLLM rollout engine.
deepspeed	zero3	DeepSpeed ZeRO Stage 3 configuration used to partition model parameters, gradients, and optimizer states.

Table 9: Training and rollout-generation hyperparameters used in the main VERA experiments.

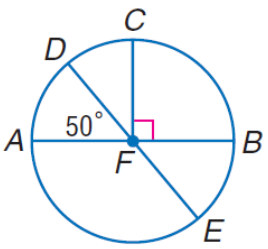
spectively. VERA should therefore be interpreted as supervising an explicit response-level evidence representation rather than enforcing a caption-first information bottleneck.

G Potential Risks

Incorrect judge feedback may reward captions that are plausible but not fully grounded or penalize valid captions that express visual relations in an unexpected form. Such errors can propagate to the reasoning policy and reduce robustness under domain shift. Language and cultural coverage is also constrained by the selected training and evaluation data.

In high-stakes settings, generated visual descriptions and answers should not be treated as verified evidence without independent review. VERA is intended for research on multimodal reasoning and does not replace domain-specific validation, uncertainty reporting, or human oversight.

Case Analysis: Circle Arc



Question: In $\odot F$, $m\angle DFA = 50°$ and $\overline{CF} \perp \overline{FB}$. Find $m\widehat{CBE}$.
Answer: $m\widehat{CBE} = 140°$.

Unfaithful and Insufficient Caption

The image depicts a circle with center F. Points A, B, C, D, and E lie on the circumference. The angle $\angle DFA$ is labeled $50°$, and a right-angle marker is shown between $\overline{CF}$ and $\overline{FB}$. The circle has radius F.

Problems: The caption incorrectly treats the center label F as a radius and omits the collinear triples A, F, B and D, F, E. It is therefore neither visually faithful nor sufficient to determine $\widehat{CBE}$.

Faithful but Insufficient Caption

The image shows a circle centered at F. Points A, B, C, D, and E lie on the circumference. Points A, F, and B are collinear. The angle $\angle DFA$ is labeled $50°$, and a right-angle marker is shown between $\overline{CF}$ and $\overline{FB}$.

Missing information: The caption omits that D, F, and E are collinear. Without this relation, the position of E and hence $\widehat{BE}$ cannot be determined.

Unfaithful but Sufficient Caption

A circle centered at F contains points A, B, C, D, and E on its circumference. Points A, F, and B lie on a horizontal line, with A to the left of F and B to the right. Points D, F, and E are collinear. Points C and D lie above the line $\overline{AB}$, whereas E lies below it. The diagram shows $\angle DFA = 50°$ and a right-angle marker between $\overline{CF}$ and $\overline{FB}$. The interior of the circle is shaded gray.

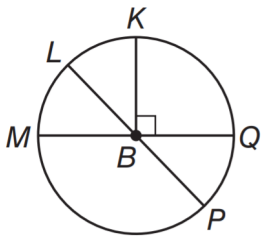
Faithfulness error: The circle is not shaded in the image. This unsupported visual claim is irrelevant to the target calculation, while the remaining information is sufficient to solve the problem.

Faithful and Sufficient Caption

The image shows a circle centered at F, with A, B, C, D, and E on the circumference. Points A, F, and B lie on a horizontal line, with A to the left of F and B to the right. Points D, F, and E are collinear. Points C and D lie above the line AB, whereas E lies below it. The angle $\angle DFA$ is labeled $50°$. A right-angle marker is shown between $\overline{CF}$ and $\overline{FB}$. The requested arc runs from C to E through B.

Figure 9: Four caption categories for the circle-arc example.

Case Analysis: Central Angles



Question: In $\odot B$, $m\angle LBM = 3x$ and $m\angle LBQ = 4x + 61$. What is $m\angle PBQ$?

Answer: $m\angle PBQ = 51°$.

Unfaithful and Insufficient Caption

The image shows a circle centered at B. Points K, B, and M are collinear, forming diameter $\overline{KM}$, while points P, B, and Q are collinear, forming diameter $\overline{PQ}$. A right-angle marker is shown between $\overline{BK}$ and $\overline{BQ}$.

Problems: The claimed diameters are not present in the image; the visible straight lines are M, B, Q and L, B, P. Because the caption also omits the actual line through L, B, and P, it cannot determine $\angle PBQ$.

Faithful but Insufficient Caption

The diagram shows a circle centered at B, with K, L, M, Q, and P on the circumference. Points M, B, and Q are collinear. A right-angle marker is shown between $\overline{BK}$ and $\overline{BQ}$. The angles $\angle LBM$, $\angle LBQ$, and $\angle PBQ$ are central angles.

Missing information: The caption omits that L, B, and P are collinear. This relation is required to transfer the measure of $\angle LBM$ to $\angle PBQ$.

Unfaithful but Sufficient Caption

The image shows a circle centered at B. Points M, B, Q are collinear, and points L, B, P are collinear. A right-angle marker is shown between $\overline{BK}$ and $\overline{BM}$.

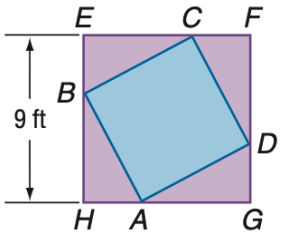
Faithfulness error: The right-angle marker is drawn between $\overline{BK}$ and $\overline{BQ}$, not between $\overline{BK}$ and $\overline{BM}$. The two correct collinear triples still provide all visual information needed to determine $\angle PBQ$.

Faithful and Sufficient Caption

The diagram shows a circle centered at B, with K, L, M, Q, and P on the circumference. Points M, B, and Q are collinear, and points L, B, and P are collinear. A right-angle marker is shown between $\overline{BK}$ and $\overline{BQ}$.

Figure 10: Four caption categories for the central-angle example.

Case Analysis: Area Ratio



Question: The vertices of quadrilateral $ABCD$ divide the sides of square $EFGH$ in a 1:2 ratio. Find $[ABCD] : [EFGH]$.

Answer: 5:9.

Unfaithful and Insufficient Caption

The square $EFGH$ has side length 9 ft. Point A divides $\overline{HG}$ with $HA : AG = 1 : 2$, and point B divides $\overline{EH}$ with $EB : BH = 1 : 2$. Point C divides $\overline{EF}$ with $EC : CF = 1 : 2$, and point D divides $\overline{FG}$ with $FD : DG = 1 : 2$.

Problems: The stated orientations for C and D contradict their visible positions; the corresponding relations are $EC : CF = 2 : 1$ and $FD : DG = 2 : 1$. The mixed orientation describes a different quadrilateral and does not yield the reference area ratio.

Faithful but Insufficient Caption

The image shows square $EFGH$ with side length 9 ft. Point A lies on $\overline{HG}$, B lies on $\overline{EH}$, C lies on $\overline{EF}$, and D lies on $\overline{FG}$. These points form the interior quadrilateral $ABCD$.

Missing information: The caption does not specify which endpoint of each side is closer to the corresponding point. The orientation of the 1:2 divisions is needed to identify the intended quadrilateral uniquely.

Unfaithful but Sufficient Caption

The square $EFGH$ has side length 9 ft. The four points satisfy $GA : AH = 1 : 2$, $HB : BE = 1 : 2$, $EC : CF = 1 : 2$, and $FD : DG = 1 : 2$.

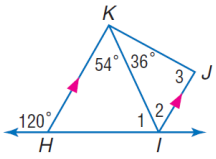
Faithfulness error: Each displayed ratio is reversed relative to the visible point placement. Reversing all four divisions produces a rotationally corresponding quadrilateral with the same area, so the caption remains sufficient for the requested area ratio.

Faithful and Sufficient Caption

The image shows square $EFGH$ with side length 9 ft. Point A lies on $\overline{HG}$ closer to H than to G; B lies on $\overline{HE}$ closer to E than to H; C lies on $\overline{EF}$ closer to F than to E; and D lies on $\overline{FG}$ closer to G than to F. The four points form quadrilateral $ABCD$.

Figure 11: Four caption categories for the area-ratio example.

Case Analysis: Parallel Lines



Question: If $\overline{KH} \parallel \overline{JI}$, find $m\angle 2$.

Answer: $54°$.

Unfaithful and Insufficient Caption

The line through H and I is horizontal. The angle between $\overline{KH}$ and $\overline{KJ}$ is labeled $54°$, and the angle between $\overline{KJ}$ and $\overline{JI}$ at J is labeled $36°$. The segment $\overline{JI}$ has length 3, and $\overline{KH} \parallel \overline{JI}$.

Problems: The $54°$ and $36°$ labels are misplaced, and the numeral 3 denotes an angle label rather than a segment length. The caption omits the actual relation $\angle HKI = 54°$, so it cannot determine $\angle 2$.

Faithful but Insufficient Caption

The diagram shows points H, I, J, and K, with H and I on a horizontal line. Segments $\overline{KH}$, $\overline{KI}$, $\overline{KJ}$, and $\overline{JI}$ are drawn. An exterior angle of $120°$ is marked at H. Angles 1 and 2 are labeled at I, and matching arrow marks indicate $\overline{KH} \parallel \overline{JI}$.

Missing information: The caption omits the marked value $\angle HKI = 54°$, which is required with the parallel-line relation to determine $\angle 2$.

Unfaithful but Sufficient Caption

The diagram shows $\overline{KH} \parallel \overline{JI}$ and $\angle HKI = 54°$. It also places a $36°$ angle at J and states that $\overline{JI}$ has length 3. Angle 2 is formed by $\overline{IK}$ and $\overline{IJ}$.

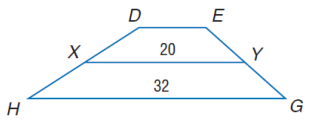
Faithfulness error: The $36°$ label is at K, not at J, and 3 is an angle label rather than a segment length. The correct $54°$ angle and the parallel-line relation are still sufficient to obtain $\angle 2$.

Faithful and Sufficient Caption

The diagram shows H and I on a horizontal line, with K above the line and J to the upper right. The angle between KH and KI at K is labeled $54°$, while the angle between KI and KJ is labeled $36°$. Angles 1 and 2 are marked at I, and angle 2 is formed by rays $\overrightarrow{IK}$ and $\overrightarrow{IJ}$. Angle 3 is marked at J. Matching arrow marks indicate $KH \parallel JI$.

Figure 12: Four caption categories for the parallel-line example.

Case Analysis: Trapezoid Midsegment



Question: For trapezoid $DEGH$, X and Y are midpoints of the legs. Find $\overline{DE}$.

Answer: $\overline{DE} = 8$.

Unfaithful and Insufficient Caption

The diagram shows trapezoid $DEGH$. The top base $\overline{DE}$ is labeled 20, the segment $\overline{XY}$ is labeled 32, and the bottom base $\overline{HG}$ is unlabeled.

Problems: The numerical labels are assigned to the wrong segments: the image shows $\overline{XY} = 20$ and $\overline{HG} = 32$. The incorrect measurements do not support the reference value of $\overline{DE}$.

Faithful but Insufficient Caption

The diagram shows trapezoid $DEGH$ with parallel bases $\overline{DE}$ and $\overline{HG}$. Points X and Y lie on the legs, and $\overline{XY}$ is parallel to the bases and labeled 20.

Missing information: The caption omits the labeled length $\overline{HG} = 32$, which is required with the midpoint condition from the problem to determine $\overline{DE}$.

Unfaithful but Sufficient Caption

The diagram shows trapezoid $DEGH$ with top base $\overline{DE}$ and bottom base $\overline{HG}$. Point X lies on leg $\overline{DH}$, and point Y lies on leg $\overline{EG}$. Segment $\overline{XY}$ is labeled 20, and the bottom base $\overline{HG}$ is labeled 32. Matching tick marks are shown on legs $\overline{DH}$ and $\overline{EG}$.

Faithfulness error: The image contains no matching tick marks on $\overline{DH}$ and $\overline{EG}$. This unsupported visual claim is irrelevant to the target calculation, while the remaining visual information, together with the midpoint condition stated in the question, is sufficient to determine $\overline{DE}$.

Faithful and Sufficient Caption

The diagram shows trapezoid $DEGH$ with top base $\overline{DE}$ and bottom base $\overline{HG}$. Point X lies on leg $\overline{DH}$ and point Y lies on leg $\overline{EG}$. Segment $\overline{XY}$ is drawn parallel to the bases and is labeled 20. The bottom base $\overline{HG}$ is labeled 32.

Figure 13: Four caption categories for the trapezoid-midsegment example.