CRAFT: TEACHING VISION-LANGUAGE REASONERS WHAT MAKES REASONING WORTH KEEPING

Anonymous authors
Paper under double-blind review

ABSTRACT

Multimodal reasoning models leverage Chain-of-Thought (CoT) reasoning to solve complex visual tasks, yet incorrect deductions, visually ungrounded descriptions, and irrelevant visual attention can degrade reasoning quality and incur unnecessary inference cost. We introduce **CRAFT** (**C**orrectness, g**R**ounding, **A**nd **F**ocus **T**raining), a discrimination-like auxiliary verification framework for multimodal CoT compression. Unlike existing approaches that directly construct, imitate, or optimize shorter reasoning chains, CRAFT shifts the training objective toward *reasoning quality assessment*. Specifically, it introduces three simple yet complementary auxiliary verification tasks targeting reasoning correctness, visual faithfulness, and visual focus. These tasks supervise the model to identify incorrect deductions, visually unsupported descriptions, and task-irrelevant visual information, thereby strengthening its internal ability to distinguish high-quality reasoning from flawed or redundant content. Through joint training, the learned verification capability transfers to open-ended multimodal reasoning, enabling the model to reduce redundant reasoning while preserving critical visual evidence, without requiring carefully constructed concise CoTs, reward models, or reinforcement learning. Experiments on four public multimodal question-answering benchmarks across multiple model scales show that CRAFT substantially reduces output token usage while maintaining answer accuracy, and further improves accuracy in most settings.

1 INTRODUCTION

Multimodal large language models (MLLMs) (Bai et al., 2025b; OpenAI et al., 2024) have made remarkable progress in visual understanding and language generation through vision-language pretraining (Liu et al., 2023). To tackle complex problems involving multi-step computation, spatial reasoning, and cross-modal evidence integration, recent multimodal large reasoning models (MLRMs) further introduce explicit Chain-of-Thought (CoT) reasoning, improving performance on challenging visual tasks through visual observation, problem decomposition, and step-by-step deduction (Xu et al., 2025a; Peng et al., 2025; Deng et al., 2025; Huang et al., 2026; Meng et al., 2025; Kimi Team et al., 2025). These advances suggest that allocating more computation to the reasoning stage can substantially enhance multimodal reasoning capabilities (Xu et al., 2025a; Kimi Team et al., 2025).

However, the performance gains brought by CoT often come with substantial reasoning redundancy and additional computational overhead (Peng et al., 2025; Zhang et al., 2026b; Xiao & Gan, 2025; Huang et al., 2026). Unlike text-only reasoning, multimodal reasoning requires not only logical deduction but also observing, describing, and selecting visual evidence relevant to the question. Rich visual inputs may induce visually unsupported descriptions or repeated attention to task-irrelevant information, further amplifying reasoning redundancy (Wang et al., 2026). Existing approaches to multimodal CoT compression mainly optimize the reasoning process from the *generation side*. One line of work constructs or filters shorter reasoning trajectories for imitation learning (Peng et al., 2025; Zhang et al., 2026b; Wang et al., 2026), while another directly constrains reasoning length through reinforcement learning (Xiao & Gan, 2025). Although these methods can reduce redundancy, they typically rely on constructing or selecting concise trajectories, or require repeated sampling, reward computation, and policy optimization, thereby introducing additional training complexity (Zhang et al., 2026b; Wang et al., 2026).

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.

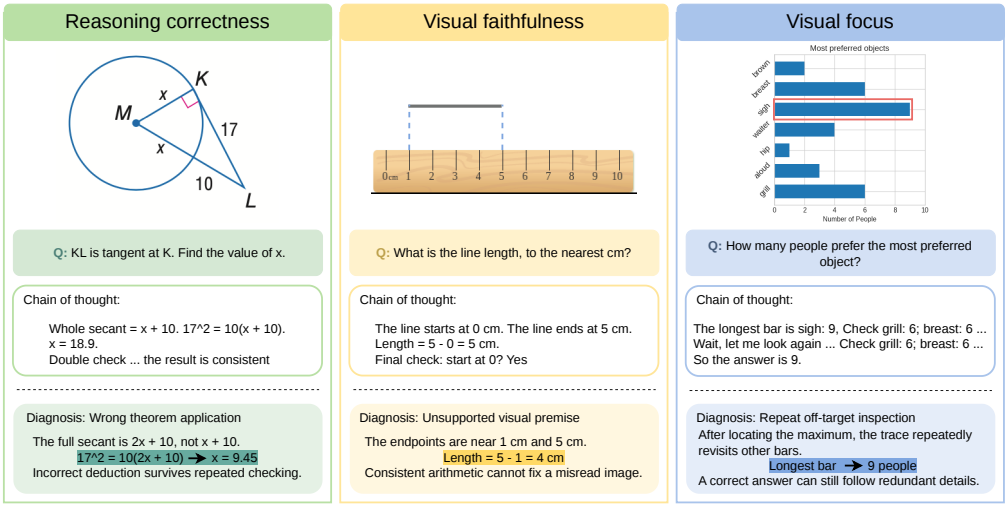


Figure 1: (green) **Reasoning correctness:** incorrect application of the tangent–secant theorem, reinforced during self-checking. (yellow) **Visual faithfulness:** an unsupported assumption that the line starts at zero leads to an incorrect measurement. (blue) **Visual focus:** redundant inspection of irrelevant bars after the correct answer has already been identified. Reasoning excerpts are shortened for clarity.

More importantly, *shorter* does not necessarily mean *better*. Relying solely on reasoning length and final-answer correctness makes it difficult to determine whether removed content is genuinely redundant or instead contains visual evidence necessary for solving the problem (Xu et al., 2025b; Wang et al., 2026). High-quality multimodal reasoning should satisfy at least three requirements: the logical derivation should be correct, visual statements should be grounded in the image, and the attended visual information should be relevant to the question. Figure 1 illustrates how incorrect deductions, unsupported visual descriptions, and irrelevant visual attention can jointly lead to unnecessarily long reasoning traces. This motivates a different question: *rather than explicitly constructing an ideal concise CoT, can we achieve more efficient multimodal reasoning simply by teaching the model to identify which reasoning steps are unreliable or unnecessary?*

To this end, we propose **CRAFT**, a multimodal CoT compression method based on discrimination-like auxiliary verification for **C**orrectness, g**R**ounding, **A**nd **F**ocus **T**raining. CRAFT reformulates reasoning quality assessment into three simple yet complementary discrimination-like auxiliary tasks: reasoning-step correctness judgment, visual faithfulness verification, and visual focus verification. Through joint training, the model learns to identify incorrect deductions, visually unsupported statements, and task-irrelevant visual information, and transfers this discriminative capability to open-ended multimodal reasoning. Importantly, these auxiliary tasks are used only during training; at inference time, CRAFT requires no additional verification, reranking, or multi-round generation, and therefore introduces no extra inference overhead. Experiments on four public multimodal question-answering benchmarks and four backbone models show that CRAFT consistently reduces output token counts across all evaluation settings while preserving accuracy on challenging problems, and even improves answer accuracy in most settings.

Our main contributions are summarized as follows:

- We construct three complementary discrimination-like auxiliary verification tasks centered on reasoning correctness, visual faithfulness, and visual focus, providing explicit supervision for incorrect deductions, visually unsupported statements, and task-irrelevant visual information.
- We propose CRAFT, which learns to assess the quality of candidate reasoning rather than directly constructing or imitating concise CoTs. Without requiring reward models or reinforcement learning, CRAFT reduces redundant reasoning while preserving necessary visual evidence.

- We evaluate CRAFT on four public multimodal question-answering benchmarks across multiple model scales. The results show that CRAFT consistently reduces output token counts while maintaining or improving answer accuracy in most settings.

2 PROBLEM SETUP

We study efficient generation for multimodal reasoning. Given an image i and a textual question q, a multimodal reasoning model M_θ solves the problem by generating a Chain-of-Thought (CoT) rationale. Our goal is to reduce unnecessary reasoning generation, and thereby inference cost, while preserving the model's problem-solving capability as much as possible. For a target task distribution $\mathcal{D}$, we formulate this objective as

$$\min_{\theta} \ \mathbb{E}_{(i,q)\sim\mathcal{D}} \left[\mathrm{Cost}(M_\theta; i, q)\right], \tag{1}$$

where $\mathrm{Cost}(M_\theta; i, q)$ denotes the inference cost for a given image–question pair, measured by the number of generated CoT tokens. Rather than simply pursuing the shortest possible output, we aim to eliminate redundant reasoning while retaining effective problem-solving capability. Therefore, efficiency gains should not come at the expense of critical reasoning steps or necessary visual evidence.

In multimodal settings, reducing generation also requires preserving the logical structure and visual evidence necessary for correct problem solving. Consequently, simply compressing reasoning steps is insufficient to characterize high-quality efficient reasoning. We therefore ask: *Can auxiliary verification training, formulated as discrimination-like tasks, teach the model to identify incorrect, visually unsupported, or question-irrelevant reasoning content, and can this assessment capability transfer to standard problem solving to reduce redundant generation while preserving problem-solving capability?*

3 METHODS

3.1 THREE-DIMENSIONAL AUXILIARY VERIFICATION FINE-TUNING

Reducing redundancy in multimodal reasoning requires models to distinguish effective reasoning from unnecessary elaboration, while preserving the visual evidence essential for solving the problem. Prior work has shown that verification-based training can improve subsequent solution generation (Chen et al., 2025; 2026). Motivated by these findings, we introduce explicit verification supervision for reasoning quality in multimodal settings. Specifically, we propose CRAFT, which trains multimodal reasoning models to assess candidate reasoning trajectories along three complementary dimensions: reasoning correctness, visual faithfulness, and visual focus. Rather than directly supervising the model to generate shorter reasoning traces, we formulate these dimensions as simple verification-style tasks and use concise natural-language feedback as supervision. This allows the model to learn to identify incorrect deductions, visually unsupported statements, and irrelevant or redundant visual observations. As illustrated in Figure 2, CRAFT jointly trains these three auxiliary verification tasks with the shared multimodal reasoner, while requiring no additional verification stage at inference time. We further investigate whether such quality-assessment capabilities can transfer to open-ended problem solving, enabling more concise reasoning while preserving problem-solving capability.

Given an image i, a question q, and a candidate reasoning trajectory r, the model generates task-specific textual feedback for a designated verification task. All three tasks share the same model parameters and are distinguished by task-specific instructions. Importantly, the category indices defined below are used only to formalize different supervision cases; the model is not trained to predict discrete numerical labels, but rather to generate the corresponding natural-language feedback.

Reasoning Correctness Verification. Incorrect deductions may lead to wrong answers and further induce unproductive reasoning branches, making correctness a fundamental requirement for effective reasoning. We therefore introduce reasoning correctness verification to teach the model to distinguish successful candidate solutions from unsuccessful ones, rather than treating reasoning length alone as an indicator of quality.

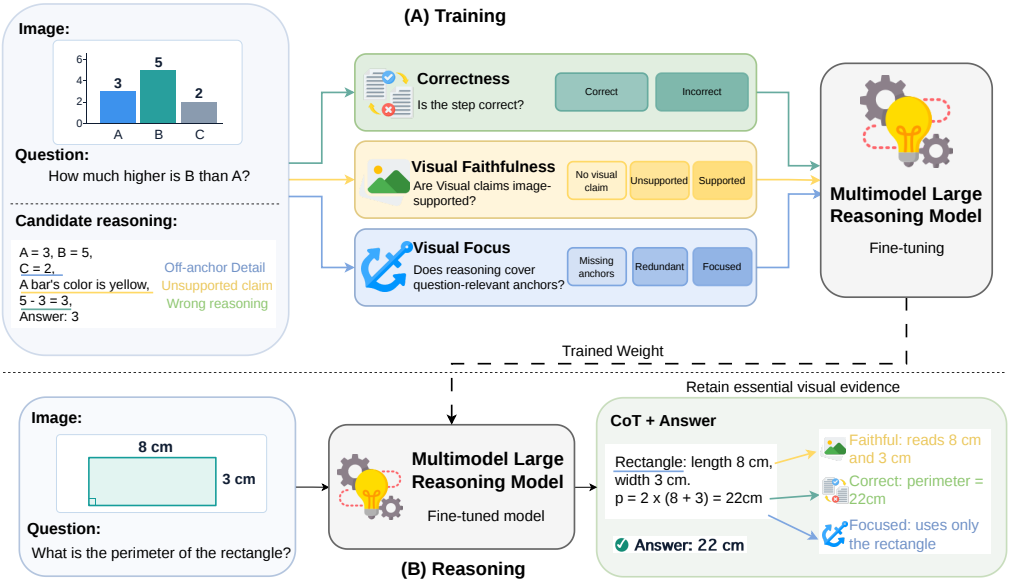


Figure 2: Overview of CRAFT. (a) During training, three auxiliary verification tasks supervise reasoning correctness, visual faithfulness, and visual focus. (b) At inference time, the model directly generates concise reasoning and answers without additional verification.

We determine correctness supervision based on agreement between the final answer and the reference answer. Let $a(r)$ denote the final answer extracted from candidate trajectory r, and let a^* denote the reference answer. We define the corresponding category as

$$z_{\text{cor}} = \begin{cases} 0, & a(r) \equiv a^*, \\ 1, & a(r) \not\equiv a^*, \end{cases} \tag{2}$$

where $\equiv$ denotes answer matching after normalization. During training, the model generates the textual feedback associated with each case rather than the numerical category itself.

The reference answer is used only to determine the supervision target, and answer agreement serves as a proxy signal for candidate reasoning correctness.

Visual Faithfulness Verification. Multimodal reasoning relies on visual observations as premises for subsequent deductions. Visually unsupported descriptions may therefore introduce incorrect or unproductive reasoning branches. Even when the final answer is correct, the reasoning process may still contain inaccurate visual descriptions, motivating visual faithfulness supervision beyond reasoning correctness.

Let $\mathcal{C}(r)$ denote the set of verifiable visual claims in candidate trajectory r, including statements about objects, attributes, spatial relations, and numerical information. Let $s(c, i) \in \{0, 1\}$ indicate whether image i supports visual claim c. We define three visual-faithfulness cases:

$$z_{\text{faith}} = \begin{cases} 0, & \mathcal{C}(r) = \varnothing, \\ 1, & \exists c \in \mathcal{C}(r) : s(c, i) = 0, \\ 2, & \mathcal{C}(r) \neq \varnothing \ \wedge \forall c \in \mathcal{C}(r) : s(c, i) = 1. \end{cases} \tag{3}$$

Again, the category indices are used only to describe different supervision cases, while the actual training target is the corresponding natural-language feedback. Separating the case with no verifiable visual description avoids conflating the absence of visual information with fully faithful visual grounding. This task therefore provides direct supervision for identifying unreliable visual descriptions.

Visual Focus Verification. Images often contain substantially more information than is needed for a particular question. As a result, even visually faithful descriptions may still introduce redundancy by attending to irrelevant content or repeatedly inspecting unnecessary details. Conversely, overly aggressive compression may discard critical evidence required for solving the problem. We therefore introduce visual focus verification to assess whether the reasoning remains centered on the visual information relevant to the question.

We refer to the visual entities and relations relevant to solving the question as *visual anchors*. Let $\mathcal{A}(q)$ denote the set of visual anchors identified from the question, including relevant objects, attributes, regions, quantities, and relations. Let $h(i, q, r) \in \{0, 1\}$ indicate whether the candidate reasoning sufficiently attends to and covers the required visual anchors, and let $u(i, q, r) \in \{0, 1\}$ indicate whether it contains irrelevant or unnecessarily repeated visual descriptions beyond these anchors. We define three visual-focus cases:

$$z_{\text{focus}} = \begin{cases} 0, & h(i,q,r) = 0, \\ 1, & h(i,q,r) = 1 \ \wedge \ u(i,q,r) = 1, \\ 2, & h(i,q,r) = 1 \ \wedge \ u(i,q,r) = 0. \end{cases} \tag{4}$$

This task jointly considers coverage of necessary visual evidence and redundancy in additional visual descriptions, thereby encouraging the model to retain critical visual anchors while reducing observations that do not contribute to solving the question.

Overall, the three verification tasks address complementary aspects of reasoning quality: whether the candidate solution is correct, whether its visual claims are supported by the image, and whether its visual observations are focused on evidence required by the question. These dimensions are not interchangeable: a trajectory with a correct final answer may still contain unsupported visual statements, while visually faithful descriptions may still be irrelevant or unnecessarily repetitive.

Joint Training Objective. Let $\mathcal{T} = \{\text{cor}, \text{faith}, \text{focus}\}$ denote the three auxiliary verification tasks. For each task $d \in \mathcal{T}$, let $\mathcal{D}_d$ denote its training distribution, s_d the corresponding task instruction, and t_d the target textual feedback determined by the rules above. Since the supervision target consists of multiple text tokens, we optimize the standard autoregressive negative log-likelihood:

$$\mathcal{L}_d(\theta) = -\mathbb{E}_{(i,q,r,t_d)\sim\mathcal{D}_d} \left[\log p_\theta(t_d \mid i, q, r, s_d) \right], \tag{5}$$

where

$$\log p_\theta(t_d \mid i, q, r, s_d) = \sum_{k=1}^{|t_d|} \log p_\theta \left(t_{d,k} \mid i, q, r, s_d, t_{d,<k} \right). \tag{6}$$

After mixing samples from all three tasks, the overall CRAFT objective is

$$\mathcal{L}_{\text{CRAFT}}(\theta) = \sum_{d\in\mathcal{T}} \mathcal{L}_d(\theta), \tag{7}$$

which can equivalently be viewed as the standard autoregressive language modeling loss over the mixed training distribution of the three verification tasks.

The auxiliary verification supervision is used only during training. At inference time, the model directly generates a solution from the image and question, without any additional verification stage, reranking procedure, or separate verifier.

3.2 From Auxiliary Verification to Efficient Generation

CRAFT uses auxiliary verification as a training signal to improve the reasoning model, rather than as a separate task that must be executed at inference time. Verification and generation share the same model parameters: the former evaluates candidate reasoning trajectories conditioned on an image and a question, while the latter directly produces a complete solution. Joint training provides explicit supervision on solution correctness, visual faithfulness, and question relevance. These signals strengthen the model's intrinsic ability to distinguish high-quality reasoning from flawed or redundant content, thereby promoting more efficient reasoning generation.

Table 1: Main results on four multimodal benchmarks. All results are averaged over multiple runs with different random seeds. Acc. denotes answer accuracy (%), and Tokens denotes the average number of output tokens. For each backbone and benchmark, the highest accuracy and lowest token count are highlighted in bold. MMMU-Pro uses the Standard setting.

	MMBench		MathVision		MathVista		MMMU-Pro	
Method	Acc.↑	Tokens↓	Acc.↑	Tokens↓	Acc.↑	Tokens↓	Acc.↑	Tokens↓
Qwen3.5-4B								
Original	87.3	721.0	45.9	6233.0	68.1	1230.7	58.3	4545.7
Prompt	88.7	678.4	**47.5**	5690.7	70.4	1089.0	57.0	3925.6
SFT	88.6	617.1	45.1	6014.9	70.0	1122.5	**58.4**	4249.2
CRAFT (ours)	**89.1**	**210.8**	46.6	**5417.7**	**72.1**	**683.6**	57.7	**3075.3**
Qwen3.5-9B								
Original	89.0	736.6	49.9	6179.6	68.7	1275.3	63.2	4563.5
Prompt	89.1	676.9	50.8	5905.1	70.9	1126.6	62.6	4077.8
SFT	89.5	635.6	**51.2**	5824.6	71.2	1126.7	**63.9**	4151.1
CRAFT (ours)	**89.9**	**554.9**	50.8	**5518.2**	**72.6**	**990.9**	63.6	**3905.3**
Qwen3-VL-2B-Thinking								
Original	**84.5**	499.0	18.2	3834.0	56.1	1210.0	24.3	1706.0
Prompt	82.9	487.5	17.9	3791.2	56.5	1109.0	25.6	1457.3
SFT	83.5	528.9	**18.7**	3732.5	54.3	1048.1	27.5	1573.5
CRAFT (ours)	83.2	**196.9**	17.7	**3079.7**	**58.3**	**593.5**	**30.5**	**1148.6**
Qwen3-VL-8B-Thinking								
Original	**90.0**	259.4	38.6	3232.5	71.6	799.0	42.4	1269.3
Prompt	89.2	248.3	38.5	3012.6	72.1	689.8	42.8	1102.3
SFT	89.9	221.2	40.7	3057.5	72.3	662.2	44.9	1153.6
CRAFT (ours)	89.7	**189.5**	**40.8**	**2578.0**	**73.9**	**545.1**	**45.2**	**594.1**

At inference time, the model receives only the image and question and directly generates a reasoning trajectory together with the final answer, without requiring additional verification labels, a separate verifier, or trajectory selection. We evaluate CRAFT under this standard problem-solving setting using answer accuracy and average output token count. Our primary objective is to determine whether auxiliary verification training can reduce generation cost while maintaining or improving answer accuracy.

4 Experiments

4.1 Experimental Setup

Models. Our backbone models are Qwen3-VL-2B-Thinking and Qwen3-VL-8B-Thinking (Bai et al., 2025a), as well as Qwen3.5-4B and Qwen3.5-9B (Qwen Team, 2026). These models are selected for their strong visual understanding and multimodal reasoning capabilities, as well as their support for explicit CoT generation. Spanning two model series and parameter scales from 2B to 9B, these backbones enable us to examine the effectiveness of CRAFT across different model sizes and capabilities.

Baselines. To investigate the role of auxiliary verification learning in multimodal CoT compression, we compare CRAFT against three representative baselines: the original model, prompt-based compression, and short-CoT SFT. Specifically, **Original** directly uses the backbone model with-

out any additional fine-tuning; **Prompt** augments the problem-solving instruction with an explicit request for concise reasoning, encouraging the model to reduce repetitive analysis and irrelevant descriptions; and **SFT** fine-tunes the model on concise reasoning trajectories. Further experimental details are provided in the appendix.

Datasets. Our evaluation covers four public multimodal benchmarks: MathVista (Lu et al., 2024), MathVision (Wang et al., 2024), MMMU-Pro (Yue et al., 2025), and MMBench (Liu et al., 2024). MathVista and MathVision focus on visually grounded mathematical reasoning, while MMMU-Pro emphasizes multidisciplinary reasoning with stronger reliance on visual evidence. MMBench provides a broader assessment of general multimodal perception and reasoning. Together, these benchmarks allow us to evaluate the effectiveness of CRAFT across diverse reasoning settings.

4.2 Experimental Results and Analysis

Overall Performance. Table 1 summarizes CRAFT across four backbones and four multimodal benchmarks. CRAFT achieves the lowest average output token count in all 16 settings, reducing tokens by 10.7%–70.8% relative to the original models while improving accuracy in 12 settings. Overall, CRAFT consistently improves the accuracy–efficiency trade-off over the original, prompt-based, and SFT baselines.

Visual Mathematical Reasoning. On MathVista, CRAFT consistently improves both accuracy and efficiency across all four backbones. Compared with the original models, it gains 2.2–4.0 accuracy points while reducing output tokens by 22.3%–51.0%. CRAFT also outperforms Prompt and SFT in all four settings: relative to Prompt, it improves accuracy by 1.7–1.8 points and reduces token usage by 12.0%–46.5%; relative to SFT, it gains 1.4–4.0 points with 12.1%–43.4% fewer tokens. On MathVision, CRAFT achieves the lowest output token count for all four backbones, reducing generation by 10.7%–20.2% relative to the original models while maintaining comparable or better accuracy in most cases. It also remains competitive against Prompt and SFT: for Qwen3-VL-8B-Thinking, CRAFT achieves 40.8% accuracy with 2578.0 tokens, compared with 38.5%/3012.6 for Prompt and 40.7%/3057.5 for SFT.

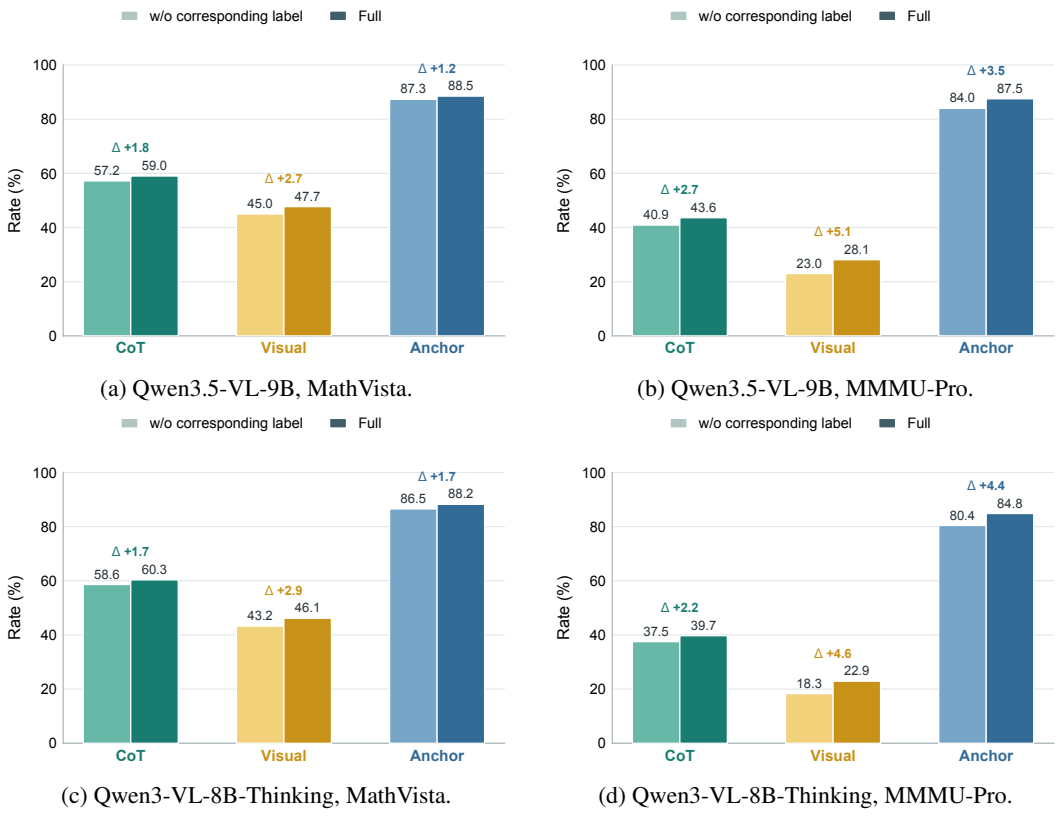
Multidisciplinary Reasoning. On MMMU-Pro, CRAFT improves the accuracy of Qwen3-VL-8B-Thinking from 42.4% to 45.2% while reducing the average output length from 1269.3 to 594.1 tokens, a reduction of 53.2%. Compared with SFT, it improves accuracy by 0.3 percentage points and further reduces output tokens by 48.5%, yielding gains in both accuracy and efficiency. Qwen3.5-9B also improves upon the original model by 0.4 percentage points in accuracy while using 14.4% fewer output tokens. Qwen3-VL-2B-Thinking achieves an accuracy gain of 6.2 percentage points. On MMBench, CRAFT reduces output token counts across all four backbones by approximately 26.9%–71.9% relative to the original models. In particular, compared with SFT, CRAFT reduces the average output length of Qwen3.5-9B from 635.6 to 562.0 tokens while increasing accuracy from 89.5% to 89.6%.

These results demonstrate the applicability of CRAFT beyond visual mathematical reasoning.

4.3 Ablation Study

We evaluate the contribution of each auxiliary verification task by comparing CRAFT with three ablated variants: *w/o CoT*, *w/o Visual*, and *w/o Anchor*, which remove correctness, visual faithfulness, and visual focus verification, respectively. As shown in Figure 3, CRAFT consistently outperforms each variant on its corresponding reasoning-quality metric across MathVista and MMMU-Pro with Qwen3.5-9B and Qwen3-VL-8B-Thinking.

Effect of Correctness Verification. Compared with *w/o CoT*, CRAFT improves the Error-Free Logical Chain Rate across both models and benchmarks. For Qwen3.5-9B, the rate increases from 57.2% to 59.0% on MathVista and from 40.9% to 43.6% on MMMU-Pro. For Qwen3-VL-8B-Thinking, it increases from 58.6% to 60.3% and from 37.5% to 39.7%, respectively. These consistent gains demonstrate the contribution of correctness verification to improving calculations and logical deductions in generated reasoning.



(a) Qwen3.5-VL-9B, MathVista.

(b) Qwen3.5-VL-9B, MMMU-Pro.

(c) Qwen3-VL-8B-Thinking, MathVista.

(d) Qwen3-VL-8B-Thinking, MMMU-Pro.

Figure 3: Ablation results on MathVista and MMMU-Pro. CoT, Visual, and Anchor denote the error-free logical-chain rate, grounded-evidence rate, and focused-anchor rate, respectively. Each pair compares removing the corresponding supervision label against the full model. Higher values are better.

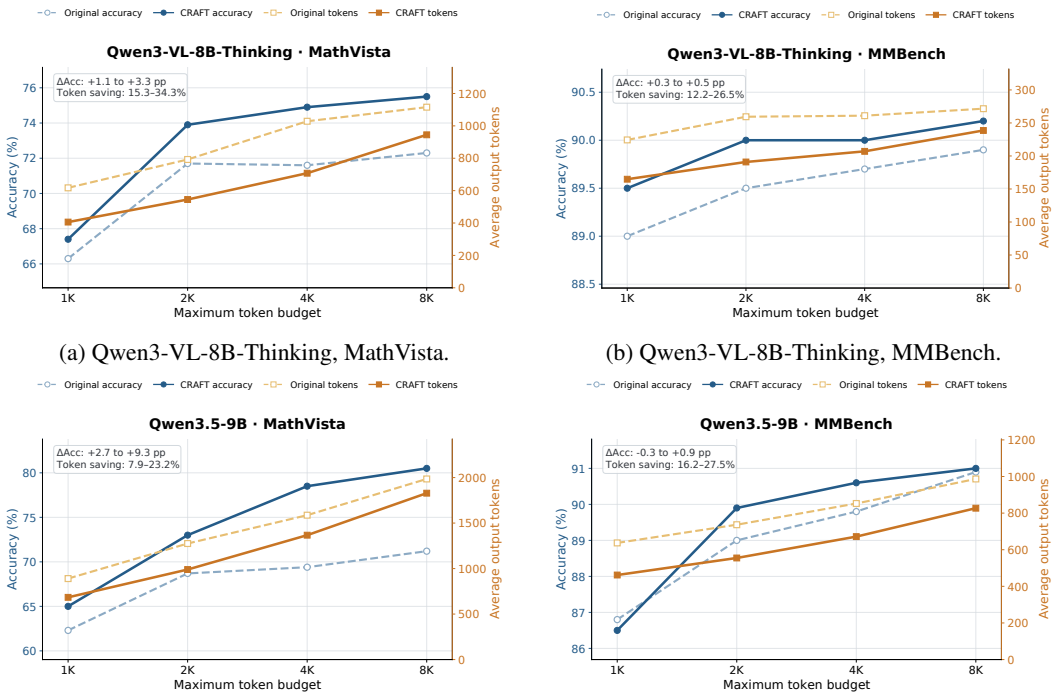
Effect of Visual Faithfulness Verification. CRAFT outperforms *w/o Visual* in Grounded Evidence Rate across all four settings. Qwen3.5-9B improves from 45.0% to 47.7% on MathVista and from 23.0% to 28.1% on MMMU-Pro. Qwen3-VL-8B-Thinking improves from 43.2% to 46.1% and from 18.3% to 22.9%, respectively. The gains of 2.7–5.1 percentage points highlight the effectiveness of visual faithfulness verification in encouraging image-supported descriptions.

Effect of Visual Focus Verification. Compared with *w/o Anchor*, CRAFT consistently improves the Focused Anchor Rate. For Qwen3.5-9B, the rate increases from 87.3% to 88.5% on MathVista and from 84.0% to 87.5% on MMMU-Pro. For Qwen3-VL-8B-Thinking, it increases from 86.5% to 88.2% and from 80.4% to 84.8%, respectively. These improvements demonstrate that visual focus verification helps models address the required visual anchors while avoiding unnecessary visual descriptions.

4.4 Effectiveness under Different Token Budgets.

We further evaluate CRAFT under different maximum generation budgets ranging from 1K to 8K tokens. As shown in Figure 4, CRAFT remains effective across a wide range of decoding budgets and consistently improves the accuracy–efficiency trade-off.

On MathVista, CRAFT improves accuracy at every token budget for both backbone models while substantially reducing the number of generated tokens. For Qwen3-VL-8B-Thinking, CRAFT yields accuracy gains of $+1.1$ to $+3.3$ percentage points together with 15.3%–34.3% token savings. The improvement is even more pronounced for Qwen3.5-9B, where accuracy increases by $+2.7$ to $+9.3$ points while output tokens are reduced by 7.9%–23.2%.



(a) Qwen3-VL-8B-Thinking, MathVista.

(b) Qwen3-VL-8B-Thinking, MMBench.

(c) Qwen3.5-9B, MathVista.

(d) Qwen3.5-9B, MMBench.

Figure 4: Performance under different maximum token budgets. CRAFT consistently improves the accuracy–efficiency trade-off across models and benchmarks.

A similar trend is observed on MMBench. For Qwen3-VL-8B-Thinking, CRAFT consistently improves accuracy by $+0.3$ to $+0.5$ points while reducing output tokens by 12.2%–26.5%. For Qwen3.5-9B, CRAFT reduces token usage by 16.2%–27.5%, while maintaining comparable or better accuracy, with changes ranging from -0.3 to $+0.9$ points. Overall, these results demonstrate that the effectiveness of CRAFT is not tied to a particular token budget. Instead, auxiliary verification training consistently encourages more economical reasoning under both constrained and relaxed generation settings.

5 Conclusion

We present **CRAFT**, a simple auxiliary verification training framework with verification-style discriminative tasks for efficient multimodal CoT reasoning. Rather than directly constructing, imitating, or optimizing shorter reasoning trajectories, CRAFT trains multimodal reasoning models to assess candidate reasoning along three complementary dimensions: reasoning correctness, visual faithfulness, and visual focus. These verification-style tasks provide explicit supervision for identifying incorrect deductions, visually unsupported descriptions, and irrelevant or redundant visual observations, enabling the learned reasoning-quality assessment capability to transfer to open-ended problem solving. Experiments on four multimodal benchmarks across multiple backbone models show that CRAFT consistently reduces output token usage while maintaining or improving answer accuracy in most settings. Its effectiveness remains robust across different generation budgets, indicating that the gains are not tied to a particular decoding length. Importantly, the auxiliary verification tasks are used only during training and require no additional verifier, reranking, or multi-stage inference at test time. These results suggest that learning to assess reasoning quality provides a simple and effective alternative to explicitly optimizing concise CoTs, offering a practical path toward more efficient multimodal reasoning.

AI USE STATEMENT

We used generative AI tools to assist with literature review, manuscript editing, data generation, and LaTeX formatting. All AI-assisted content was reviewed by the authors. Suggestions generated by AI were used only as aids for drafting and discussion, and were independently evaluated before being incorporated into the work. We take full responsibility for the final content of this work.

REPRODUCIBILITY STATEMENT

We provide detailed information throughout the main paper and appendix to support reproducibility. Section A.1 describes the training data construction pipeline, including source sampling, candidate generation, three-dimensional verification supervision, disagreement-based filtering, solver replay, and the final training mixture. Section 3.1 presents the construction of the three auxiliary verification tasks and the joint training objective, together with the corresponding supervision cases and natural-language feedback design. We also report the backbone models, evaluation benchmarks, decoding settings, and evaluation metrics used in our experiments. The appendix further provides detailed annotation prompts, training instance formats, and other implementation details. These materials are intended to provide sufficient information to reproduce the data construction, training procedure, and evaluation protocol used in this work. Upon acceptance, we will also release the code, processed training data, and model weights to further facilitate reproducibility.

REFERENCES

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-VL technical report, 2025a. URL https://arxiv.org/abs/2511.21631.

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, 2025b. URL https://arxiv.org/abs/2502.13923.

Yuxin Chen, Yu Wang, Yi Zhang, Ziang Ye, Zhengzhou Cai, Yaorui Shi, Qi Gu, Hui Su, Xunliang Cai, Xiang Wang, An Zhang, and Tat-Seng Chua. Learning to self-verify makes language models better reasoners. In *Proceedings of the 43rd International Conference on Machine Learning*, 2026. URL https://arxiv.org/abs/2602.07594.

Zigeng Chen, Xinyin Ma, Gongfan Fang, Ruonan Yu, and Xinchao Wang. VeriThinker: Learning to verify makes reasoning model efficient. In *Advances in Neural Information Processing Systems*, volume 38, 2025. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/bcdaaa1aec3ae2aa39542acefdec4e4b-Abstract-Conference.html.

Yihe Deng, Hritik Bansal, Fan Yin, Nanyun Peng, Wei Wang, and Kai-Wei Chang. OpenVLThinker: Complex vision-language reasoning via iterative SFT-RL cycles. In *Advances in Neural Information Processing Systems*, volume 38, 2025. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/b33286edd8602ead8cb966358b01d115-Abstract-Conference.html.

Minjie Hong, Zirun Guo, Jiabao Zhang, Zehan Wang, Ziang Zhang, Tao Jin, and Zhou Zhao. View-R1: Asymmetric policy optimization for difficulty-aware multimodal reinforcement learning. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 30791–30803. Association for Computational Linguistics, 2026. doi: 10.18653/v1/2026.findings-acl.1538. URL https://aclanthology.org/2026.findings-acl.1538/.

Wenxuan Huang, Bohan Jia, Shaosheng Cao, Zheyu Ye, Fei Zhao, Zhe Xu, Yao Hu, and Shaohui Lin. Vision-R1: Incentivizing reasoning capability in multimodal large language models. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/680b2a8135b9c71278a09cafb605869e-Abstract-Conference.html.

Kimi Team, Angang Du, Bofei Gao, Bowei Xing, Changjiu Jiang, Cheng Chen, Cheng Li, Chenjun Xiao, Chenzhuang Du, Chonghua Liao, et al. Kimi k1.5: Scaling reinforcement learning with LLMs, 2025. URL https://arxiv.org/abs/2501.12599.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *Advances in Neural Information Processing Systems*, volume 36, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/6dcf277ea32ce3288914faf369fe6de0-Abstract-Conference.html.

Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, Kai Chen, and Dahua Lin. MMBench: Is your multi-modal model an all-around player? In *European Conference on Computer Vision*, pp. 216–233. Springer, 2024. doi: 10.1007/978-3-031-72658-3_13. URL https://www.ecva.net/papers/eccv_2024/papers_ECCV/html/959_ECCV_2024_paper.php.

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

Fanqing Meng, Lingxiao Du, Zongkai Liu, Zhixiang Zhou, Quanfeng Lu, Daocheng Fu, Tiancheng Han, Botian Shi, Wenhai Wang, Junjun He, Kaipeng Zhang, Ping Luo, Yu Qiao, Qiaosheng Zhang, and Wenqi Shao. MM-Eureka: Exploring the frontiers of multimodal reasoning with rule-based reinforcement learning, 2025. URL https://arxiv.org/abs/2503.07365.

OpenAI et al. GPT-4o system card, 2024. URL https://arxiv.org/abs/2410.21276.

Yi Peng, Peiyu Wang, Xiaokun Wang, Yichen Wei, Jiangbo Pei, Weijie Qiu, Ai Jian, Yunzhuo Hao, Jiachun Pan, Tianyidan Xie, Li Ge, Rongxian Zhuang, Xuchen Song, Yang Liu, and Yahui Zhou. Skywork R1V: Pioneering multimodal reasoning with chain-of-thought, 2025. URL https://arxiv.org/abs/2504.05599.

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

Wenhao Shi, Zhiqiang Hu, Yi Bin, Junhua Liu, Yang Yang, See-Kiong Ng, Lidong Bing, and Roy Ka-Wei Lee. Math-LLaVA: Bootstrapping mathematical reasoning for multimodal large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.findings-emnlp.268. URL https://aclanthology.org/2024.findings-emnlp.268/.

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. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/ad0edc7d5fa1a783f063646968b7315b-Abstract-Datasets_and_Benchmarks_Track.html.

Yizhi Wang, Linan Yue, and Min-Ling Zhang. Bridging efficiency and transparency: Explainable CoT compression in multimodal large reasoning models, 2026. URL https://arxiv.org/abs/2602.09485.

Wenyi Xiao and Leilei Gan. Fast-slow thinking GRPO for large vision-language model reasoning. In *Advances in Neural Information Processing Systems*, volume 38, 2025. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/fa9259a773e85c49624aea9de2ff9146-Abstract-Conference.html.

Guowei Xu, Peng Jin, Ziang Wu, Hao Li, Yibing Song, Lichao Sun, and Li Yuan. LLaVA-CoT: Let vision language models reason step-by-step. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 2087–2098, 2025a. URL https://openaccess.thecvf.com/content/ICCV2025/html/Xu_LLaVA-CoT_Let_Vision_Language_Models_Reason_Step-by-Step_ICCV_2025_paper.html.

Zhongxing Xu, Chengzhi Liu, Qingyue Wei, Juncheng Wu, James Y Zou, Xin Wang, Yuyin Zhou, and Sheng Liu. More thinking, less seeing? assessing amplified hallucination in multimodal reasoning models. In *Advances in Neural Information Processing Systems*, volume 38, 2025b. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/777db387a5ccb131ba8c7cd155166b85-Abstract-Conference.html.

Guangheng Yang, Zhenliang Ni, Zhenkai Wu, Han Shu, Juan Feng, Wenming Yang, and Jie Hu. MCPO: Modality-contrastive preference optimization for multimodal chain-of-thought compression, 2026. URL https://arxiv.org/abs/2609.04947.

Xiang Yue, Tianyu Zheng, Yuansheng Ni, Yubo Wang, Kai Zhang, Shengbang Tong, Yuxuan Sun, Botao Yu, Ge Zhang, Huan Sun, Yu Su, Wenhu Chen, and Graham Neubig. MMMU-Pro: A more robust multi-discipline multimodal understanding benchmark. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 15134–15186. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.acl-long.736. URL https://aclanthology.org/2025.acl-long.736/.

Dongxu Zhang, Yiding Sun, Cheng Tan, Wenbiao Yan, Ning Yang, Jihua Zhu, and Haijun Zhang. Chain-of-thought compression should not be blind: V-Skip for efficient multimodal reasoning via dual-path anchoring. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 19402–19415. Association for Computational Linguistics, 2026a. doi: 10.18653/v1/2026.acl-long.887. URL https://aclanthology.org/2026.acl-long.887/.

Ruiyang Zhang, Dongzhan Zhou, and Zhedong Zheng. SketchThinker-R1: Towards efficient sketch-style reasoning in large multimodal models. In *The Fourteenth International Conference on Learning Representations*, 2026b. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/a1f0c0cd6caaa4863af5f12608edf63e-Abstract-Conference.html.

A Appendix

A.1 Training Data Construction

A.1.1 Source Sampling and Candidate Generation

We construct our training corpus from MathV360K (Shi et al., 2024) and LLaVA-CoT (Xu et al., 2025a). Specifically, we sample 15,000 image–question pairs from MathV360K. For each sample, we collect candidate reasoning trajectories from three vision–language models: Qwen3.5-4B, Qwen3-VL-8B-Instruct, and Qwen3.5-9B. Using heterogeneous generators exposes model-specific reasoning failure modes and, more importantly, provides diverse candidate trajectories for the same visual problem.

During candidate generation, we explicitly encourage concise reasoning. Given an image i and a question q, each generator receives the following prompt:

> **System:** You are a helpful multimodal reasoning assistant.
>
> **User:** {question} Solve the problem as concisely as possible. Include only the necessary concrete and verifiable visual observations that directly support the solution. Do not include unsupported, speculative, repetitive, or unrelated details. Put the final answer in \(\boxed{...}\).

Here, a concise reasoning trajectory is not obtained by mechanically truncating a longer response. Instead, it is expected to preserve the minimal visual evidence and logical derivation necessary to reach the final answer. This construction therefore avoids removing intermediate steps that are still essential for verifiable problem solving.

A.1.2 THREE-DIMENSIONAL PROCESS SUPERVISION

Let $r_q^{(m)}$ denote the reasoning trajectory generated by model m for question q. We annotate each trajectory along three complementary dimensions: reasoning correctness, visual faithfulness, and visual focus.

Reasoning Correctness. We extract the final answer from the model output and use answer equivalence as a deterministic proxy for reasoning correctness. If the predicted answer matches the reference answer, the corresponding supervision text is *"Every step appears to be absolutely correct."*; otherwise, the target is *"Some mistakes may exist in the proposed solution."*

Visual Faithfulness. A multimodal evaluator receives $(i, q, r_q^{(m)})$ and first determines whether the reasoning trajectory contains concrete and verifiable visual claims. If so, it further checks whether each claim is supported by the image. We define three supervision cases: (i) no verifiable visual description is present; (ii) at least one visual description is unsupported by the image; and (iii) the trajectory contains a non-empty set of visual descriptions, all of which are supported by the image. A single unsupported or hallucinated visual claim is sufficient to assign the second case.

Visual Focus. We first extract a set of visual anchors A_q from the question, including the entities, attributes, regions, values, and relations that must be visually inspected to answer the question. A text-only evaluator then receives $(q, A_q, r_q^{(m)})$ and assigns one of three supervision cases: missing essential visual anchors; covering the required anchors but introducing redundant visual descriptions; or properly focusing on the required visual anchors. This dimension evaluates whether the reasoning focuses on question-relevant visual evidence, while the factual validity of visual claims is handled independently by visual faithfulness.

A.1.3 DISAGREEMENT-BASED SAMPLE SELECTION

If all generators assign the same label to a problem, the resulting examples provide limited information for learning differences in reasoning quality across candidate trajectories. Therefore, for verification task k, we retain only questions for which the generators disagree on the corresponding supervision label. Formally,

$$\mathcal{D}_k = \bigcup_{q:\,|\{\ell_k(r_q^{(m)})\}_{m=1}^{M}|>1} \left\{ \left(i_q, q, r_q^{(m)}, \ell_k(r_q^{(m)})\right) \right\}_{m=1}^{M}, \tag{8}$$

where $M = 3$ and $k \in \{\text{COR}, \text{FAITH}, \text{FOCUS}\}$. The filter is applied independently to each supervision dimension, so the same problem may be retained for one verification task but excluded from another. This procedure yields 8,121 correctness-verification examples from 2,707 disagreement problems, 11,772 visual-faithfulness examples from 3,925 problems, and 15,611 visual-focus examples from 5,205 problems, resulting in 35,504 verification records in total.

A.1.4 SHORT-REASONING REPLAY AND FINAL TRAINING MIXTURE

Training only on verification objectives may weaken the model's ability to generate complete solutions. We therefore introduce a solver-replay objective that directly maps (i, q) to a concise generated solution. We filter the candidate solutions by removing empty and failed generations, and construct 53,256 solver examples. Merging the solver and verifier records results in 88,760 MathV360K training instances.

In addition, we sample 20,000 LLaVA-CoT records and construct the corresponding three-dimensional verification data in the same manner. The final training set contains 108,760 training instances in total.

A.1.5 TRAINING INSTANCE FORMAT

Each verification instance contains an image, an explicit task instruction, the original question, and a candidate reasoning trajectory. Visual-focus verification instances additionally include the extracted visual-anchor list. The supervision target is a natural-language label. In contrast, solver instances

use the generated concise reasoning trajectory as the target. As a result, all four objectives share the same multimodal conversational interface and can be directly mixed during supervised fine-tuning.

A.1.6 Annotation Prompt Templates

Below, we provide the prompt templates used for annotation. Evaluators are instructed to return exactly one label without any free-form explanation, thereby avoiding ambiguity during parsing.

Visual-Faithfulness Evaluator.

> **System:** You are a strict visual-faithfulness evaluator. Determine whether the reasoning chain contains concrete, verifiable visual descriptions and whether every visual description is fully supported by the provided image. Apply the rules in order: (1) return *"No verifiable visual description appears in the reasoning chain."* if no concrete image-checkable claim is present; (2) otherwise return *"The image does not support the visual description provided in the reasoning chain."* if any visual claim is incorrect, contradicted, unverifiable, speculative, or hallucinated; (3) otherwise return *"Evidence from the image supports the visual description in the reasoning chain."* Evaluate every visual claim and return only the exact label sentence.
>
> **User:** `<question> {question} </question> <solution> {reasoning chain} </solution>`

Visual-Anchor Extractor.

> **System:** You are a strict visual-anchor extraction system. Extract only the visual anchors explicitly mentioned or directly required by the question. An anchor may be an image-grounded entity, property, region, value, or relation that must be visually inspected. Use short, specific noun phrases; remove duplicates; do not infer extra scene details or solve the question. Return exactly one line in the form `Anchors: short anchor 1, short anchor 2, ...`. If no anchor is identifiable, return `Anchors: none`.
>
> **User:** `<question> {question} </question>`

Visual-Focus Evaluator.

> **System:** You are a strict evaluator of visual focus in multimodal reasoning. Apply the rules in order: (1) return *"Insufficient attention is given to the visual anchors mentioned in the text."* if any essential anchor is ignored, the reasoning chain addresses a different target, or the discussion is too vague; (2) otherwise return *"Redundant visual descriptions beyond the text-mentioned anchors are included in the reasoning chain."* if unnecessary visual details are introduced; (3) otherwise return *"Proper focus is placed on the visual anchors mentioned in the text."* Judge topical focus only, not the factual correctness of a visual claim, and return only the exact label sentence.
>
> **User:** `<question> {question} </question> <anchors> {anchors} </anchors> <reasoning_chain> {reasoning chain} </reasoning_chain>`

A.2 Experimental Details.

Supervised fine-tuning is performed with LoRA on a single NVIDIA RTX PRO 6000 GPU. LoRA is applied to the query, key, value, and output projections of the attention modules, with rank $r = 64$, scaling parameter $\alpha = 128$, and dropout of 0.05. We train with the negative log-likelihood loss and the AdamW optimizer, using a learning rate of 3×10^{-6}, weight decay of 0.01, and a cosine learning rate schedule with 60 warmup steps. The maximum sequence length is set to 8,192 tokens. We use BF16 precision, clip the gradient norm at 1.0, and set the random seed to 42.

All models are evaluated using vLLM on a single GPU, with a maximum context length of 16,384 tokens and a maximum generation length of 4,096 tokens per response. Decoding uses sampling with a temperature of 0.6 and top-p of 0.95, a random seed of 42.

A.3 Related Works

A.3.1 CoT Compression through Generation Optimization

CoT compression aims to reduce the cost of intermediate reasoning while preserving the information required for task completion. Existing methods typically achieve this by constructing concise reasoning trajectories or optimizing generation policies. In multimodal settings, Skywork R1V (Peng et al., 2025) adopts adaptive-length CoT distillation to mitigate excessive reasoning. SketchThinker-R1 (Zhang et al., 2026b) rewrites long reasoning traces into concise, sketch-style trajectories that retain essential logical steps. It first establishes this reasoning style through supervised fine-tuning and subsequently reinforces it through reinforcement learning guided by the SketchJudge reward model. These approaches encourage more compact reasoning through concise demonstrations and explicit supervision of reasoning style.

Another line of work regulates reasoning length through rewards and regularization. FAST-GRPO (Xiao & Gan, 2025) incorporates adaptive length rewards and difficulty-aware KL regularization to allocate reasoning budgets according to question difficulty. APO (Hong et al., 2026) applies asymmetric optimization to correct and incorrect trajectories, using Suboptimal Trajectory Complexity Regularization to penalize overly long incorrect responses and mitigate overthinking.

A.3.2 Visual Evidence-Aware CoT Compression

Visual descriptions in multimodal CoT may contain redundant information, but they can also provide essential links between image content and subsequent deductions. Compression must therefore account for the preservation of visual evidence. V-Skip (Zhang et al., 2026a) combines linguistic surprisal with cross-modal attention signals to assess token importance, reducing the risk of discarding critical visual anchors. XMCC (Wang et al., 2026) trains an explainable multimodal compressor through reinforcement learning. Its reward combines format compliance, answer correctness, stepwise criticality, and a length penalty to produce concise trajectories, which are subsequently used to fine-tune the reasoning model. MCPO (Yang et al., 2026) compares reasoning under image-conditioned and image-ablated inputs to identify visually independent steps. It combines step-level pruning with modality-contrastive preference optimization to shorten reasoning while preserving visual dependence.