BIFOCAL SELF-REWARDING: ALIGNING ANSWER SUFFICIENCY WITH VISUAL FIDELITY IN VLM REASONING

Anonymous authors
Paper under double-blind review

ABSTRACT

Outcome-reward reinforcement learning (RL) can improve multimodal reasoning without establishing whether the reasoning used to produce an answer is both sufficient for that answer and faithful to the image. Existing methods address complementary aspects of this problem: some verify the connection between intermediate reasoning and the final answer, while others strengthen the extraction or use of visual evidence. Yet neither direction alone establishes whether the same reasoning chain satisfies both properties. We introduce Bifocal Self-Rewarding, a dual-view framework that evaluates the same reasoning chain along these two dimensions. In the image-stripped view, the current policy uses its own reasoning chain to recover the answer without access to the image, testing whether the chain itself contains sufficient information. In the image-conditioned view, the frozen RL initialization provides a fixed perceptual reference, favoring correct chains that contain concrete visual evidence supported by the image. Using the current policy for sufficiency allows the evaluator to follow changes in the policy's reasoning, whereas freezing the fidelity evaluator provides a fixed reference throughout training. Both rewards are therefore assigned directly to the reasoning chain used to produce the answer, without requiring annotated visual claims or a separately trained judge. Across diverse benchmarks spanning reasoning, perception, extraction, and hallucination, Bifocal consistently improves multimodal reasoning across Qwen2.5-VL and Qwen3-VL backbones from 3B to 8B parameters while preserving visual capabilities.

1 INTRODUCTION

Reinforcement learning (RL) improves the reasoning of language models (Guo et al., 2025) and vision-language models (VLMs) (Huang et al., 2026b), but a correct answer need not rest on correct visual evidence. RL-trained VLMs can produce fluent chains of thought (CoTs) that cite details absent from the image (Xia et al., 2026; Huang et al., 2026a; Zhang et al., 2026). A binary correctness reward gives such trajectories the same credit as visually grounded ones, allowing linguistic shortcuts to persist.

Two directions address this failure. Self-verification recovers answers from descriptions without viewing the image (Li et al., 2026; Xia et al., 2026); perception-enhanced RL adds visual supervision through annotations, scorers, or auxiliary tasks. Yet recovery does not establish visual support, and stronger perception does not establish whether the answer-producing chain carries sufficient evidence. Caption coevolution improves a separate description's utility (Miao et al., 2026), leaving a mismatch between the evaluated and answer-producing objects.

We distinguish *answer sufficiency*, whether the chain enables image-free answering given the question and options, from *visual fidelity*, whether the image supports its visual claims. Accurate observations can be incomplete; a self-contained chain can rely on invented evidence. Both properties must be assessed on the same answer-producing chain.

Bifocal Self-Rewarding applies these two views to each sampled chain (Figure 2). The current policy π_θ re-reasons without the image to test answer sufficiency. Its frozen RL initialization π_0 retains the image and favors correct trajectories with concrete, image-checkable evidence. Only π_θ is trained:

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.

its verifier follows the evolving reasoning language, while π_0 preserves a stable perceptual reference with the same initial visual and linguistic interface. Both signals shape one GRPO update, without a separate description layer, annotated visual claims, or an external training evaluator.

We curate *Bifocal-24K*, a 19-source mixture of science, visual understanding, and multimodal mathematics under one multiple-choice interface. We evaluate Bifocal across 15 benchmarks and four Qwen2.5-VL/Qwen3-VL backbones from 3B to 8B, examining reasoning together with visual extraction and hallucination. Component ablations, paired interventions, and output-chain audits examine how answer recovery and visual anchoring contribute to these gains.

2 Related Work

Outcome and self-verification rewards. GRPO (Shao et al., 2024) compares sampled responses through group-relative rewards, but outcome-only feedback cannot distinguish correct answers supported by different visual evidence. Vision-R1 (Huang et al., 2026b), VL-Rethinker (Wang et al., 2025), and MM-Eureka (Meng et al., 2026) advance multimodal RL while primarily rewarding final answers. Vision-SR1 (Li et al., 2026) and Visionary-R1 (Xia et al., 2026) additionally verify generated descriptions through image-free recovery. PRCO (Miao et al., 2026) rewards evidence captions by downstream Solver success, with an auxiliary LLM suppressing answer leakage. These three methods assess an intermediate description. ACRE (Huang et al., 2025) instead verifies the think chain under option perturbations but retains the image, allowing a direct visual route to the answer.

Visual-evidence supervision. Perception-R1 (Xiao et al., 2026) uses a text-only LLM judge to assess responses against teacher-derived visual annotations; PeBR-R1 (Chen et al., 2025b) rewards descriptions using FGCLIP and teacher keywords. Perceval (Min et al., 2026) detects token-level perceptual errors with a trained process reward model, PEARL (Zhang et al., 2026) uses perception checklists to reward perception and gate reasoning, and VPPO (Huang et al., 2026a) derives visual dependence from image masking. Bifocal couples image-stripped recovery with the policy's own frozen visual prior on the same answer-producing chain (extended comparison in Appendix F).

3 Problem Formulation

For each query $(V, Q, \mathcal{C})$, π_θ generates reasoning $\mathbf{o}$ and answer a. The outcome reward $R_{\text{acc}} = \mathbf{1}[a = y^*]$ credits all sampled tokens under GRPO (Eq. 6), but cannot distinguish grounded from shortcut-based chains with equally correct answers. We classify feedback by answer sufficiency and visual evidence. For scalar-reward GRPO, a schematic representation is

$$A = \text{Norm}(R), \qquad R = R_{\text{acc}} + \mathbf{1}_{[\text{suf}]} R_{\text{suf}} + \mathbf{1}_{[\text{fid}]} R_{\text{fid}} + R_{\text{reg}}, \tag{1}$$

The indicators mark sufficiency verification over $\mathbf{o} \in \{\mathbf{t}, \mathbf{d}\}$ (answer chain or auxiliary description) and visual-evidence assessment. This taxonomy (Figure 1) identifies information sources, not every method's exact reward or update rule.

Two properties of a reasoning chain. A chain is *sufficient* if it contains enough information to determine the answer given the question and options, and visually *faithful* if the image supports its visual claims:

$$\mathcal{S}(\mathbf{o}) = \Pr(\tilde{a} = y^* \mid Q, \mathcal{C}, \mathbf{o}), \qquad \mathcal{F}(\mathbf{o}, V) = \text{Support}(\text{Claims}_{\text{vis}}(\mathbf{o}), V). \tag{2}$$

A self-contained chain can rely on invented evidence, while accurate observations can be insufficient to solve the task. Bifocal tests both properties on the same answer-producing chain.

3.1 The Sufficiency Dimension

Recovering an answer from an intermediate trace tests sufficiency but leaves two distinct gaps: which text is verified, and whether its visual claims are supported.

Proxy gap. Vision-SR1 and Visionary-R1 verify the separately generated description $\mathbf{d}$, while the answer is produced by the think chain that conditions on it,

$$\begin{aligned} &\mathbf{d} \sim \pi_\theta(\cdot \mid V, Q), \quad \mathbf{t} \sim \pi_\theta(\cdot \mid V, Q, \mathbf{d}), \\ &a = \arg\max_{c \in \mathcal{C}} \pi_\theta(c \mid V, Q, \mathbf{d}, \mathbf{t}), \end{aligned} \tag{3}$$

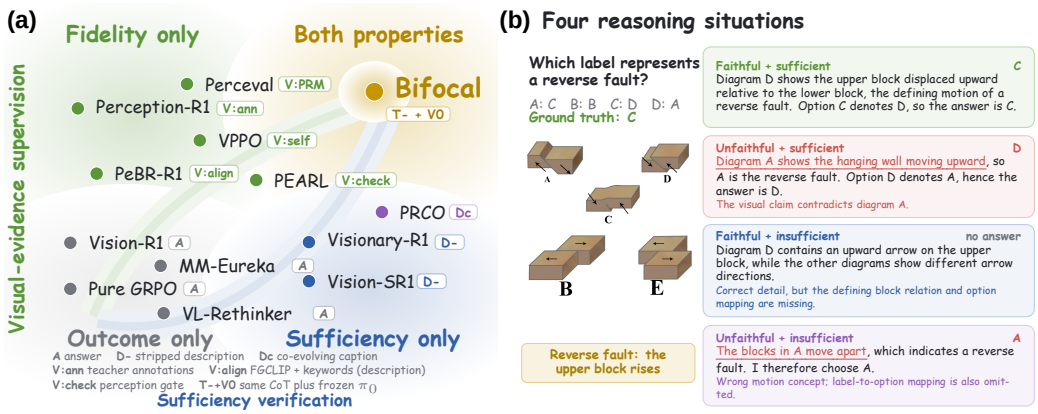


Figure 1: Two-axis method taxonomy and four reasoning states. (a) Proximity denotes methodological similarity; tags mark verified objects or visual-supervision sources. Paths summarize research directions, not direct lineage. (b) One AI2D item illustrates faithful but insufficient and sufficient but unfaithful traces; only their intersection meets both criteria.

Verifying $\mathbf{d}$ therefore constrains only $\mathcal{S}(\mathbf{d})$. Unless the answer-producing $\mathbf{t}$ retains and uses the relevant information,

$$\mathcal{S}(\mathbf{d}) \text{ high } \not\Rightarrow \mathcal{S}(\mathbf{t}) \text{ high.} \tag{4}$$

The downstream chain may omit or misread a sufficient description, or introduce unsupported claims. Verifying $\mathbf{d}$ therefore leaves the trajectory that determines a uncertified.

Blindness. Write $R_{\text{suf}}(\mathbf{o}, a, y^*; V)$ for the reward induced by a chain originating from V. Hold $(Q, \mathcal{C}, \mathbf{o}, a, y^*)$ fixed while replacing V with V'. With the same greedy decoder, the image-stripped verifier receives identical inputs and gives the same reward:

$$R_{\text{suf}}(\mathbf{o}, a, y^*; V) = R_{\text{suf}}(\mathbf{o}, a, y^*; V') \quad \text{even when} \quad \mathcal{F}(\mathbf{o}, V) \neq \mathcal{F}(\mathbf{o}, V'). \tag{5}$$

Text-only recovery therefore cannot identify visual fidelity. We use the current policy as verifier to follow its evolving CoT vocabulary and reasoning conventions, rather than an earlier trace distribution.

3.2 The Fidelity Dimension

We initialize π_θ from π_0 and update only π_θ. Reasoning RL can amplify linguistic shortcuts alongside rewarded reasoning patterns; the frozen initialization retains the model's original visual preference, unaffected by those updates, as a *perceptual anchor*.

The two tests require opposite image access. After generating $(\mathbf{t}, a)$ from $(V, Q, \mathcal{C})$, sufficiency verification removes only V and predicts $\tilde{a}$ from $(Q, \mathcal{C}, \mathbf{t})$. This blocks direct image-to-answer recovery while preserving task context. Image-derived evidence must pass through $\mathbf{t}$, although the question and options can still aid recovery. Fidelity assessment instead retains V, because visual support relates the chain's claims to the image. Both tests therefore target the same chain through complementary information interventions.

The observer shares the policy's initial visual and linguistic interface. Its preference $\widehat{\mathcal{F}}_{\pi_0}(\mathbf{t}, V, Q)$ favors concrete, image-checkable evidence among correct rollouts for the same query. Group-relative scoring requires no calibration across prompts or backbones.

The shared pre-RL prior serves two distinct roles. KL constrains overall policy drift but does not rank equally successful chains by their visual evidence; the frozen preference supplies that comparison. The current verifier follows changing reasoning conventions, while the observer provides a fixed visual reference. Ablations test these roles (Section 5.3), and paired interventions probe their signals (Section 5.4).

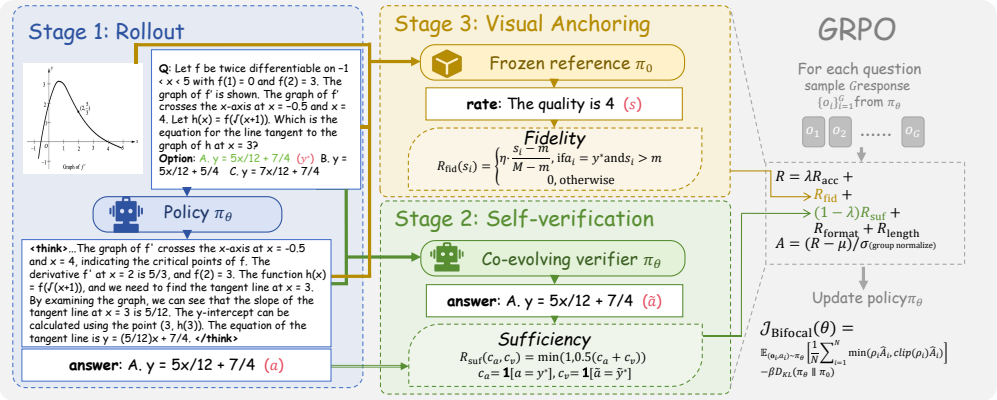


Figure 2: Overview of Bifocal Self-Rewarding. The current policy tests answer sufficiency through image-stripped recovery (green), while its frozen RL initialization assesses visual evidence with the image (orange). Both signals enter one GRPO update.

4 Bifocal Self-Rewarding

4.1 Two-View Policy Update

GRPO (Shao et al., 2024) uses within-group standardized rewards instead of a learned critic. For query $x = (V, Q, \mathcal{C})$, sample $G(x) = \{(\mathbf{t}_i, a_i)\}_{i=1}^N$ from $\pi_{\theta_{\text{old}}}$. Let $\rho_i = \pi_\theta(\mathbf{t}_i, a_i \mid x)/\pi_{\theta_{\text{old}}}(\mathbf{t}_i, a_i \mid x)$ and $\bar{\rho}_i = \text{clip}(\rho_i, 1-\epsilon, 1+\epsilon)$. Bifocal maximizes

$$\mathcal{J}_{\text{Bifocal}}(\theta) = \mathbb{E}_{x,G\sim\pi_{\theta_{\text{old}}}}\left[\frac{1}{N}\sum_{i=1}^{N} \min\left(\rho_i \widehat{A}_i^{\text{Bifocal}}, \bar{\rho}_i \widehat{A}_i^{\text{Bifocal}}\right) - \beta D_{\text{KL}}(\pi_\theta \,\|\, \pi_0)\right]. \tag{6}$$

Both views assess the same $\mathbf{t}_i$, yielding

$$\widehat{A}_i^{\text{Bifocal}} = \text{Norm}_{G(x)}\Bigg(\underbrace{\lambda R_{\text{acc},i} + (1-\lambda) R_{\text{suf}}(\mathbf{t}_i; Q, \mathcal{C})}_{\substack{\text{answer-carrying view} \\ \text{(image stripped)}}} + \underbrace{R_{\text{fid}}(\mathbf{t}_i; V, Q, \pi_0)}_{\substack{\text{visual-anchor view} \\ \text{(image retained)}}} + R_{\text{format},i} + R_{\text{length},i}\Bigg). \tag{7}$$

Here λ balances answer correctness and sufficiency, and $\text{Norm}_{G(x)}(R_i) = (R_i - \mu_G)/(\sigma_G + \delta)$ uses group statistics and stabilizer δ. Both passes supply detached rewards; only π_θ is updated. The objective uses compact trajectory-ratio notation.

4.2 Sufficiency Certification

For $(\mathbf{t}, a) \sim \pi_\theta(\cdot \mid V, Q, \mathcal{C})$, the current policy recovers $\tilde{a}$ from $(Q, \mathcal{C}, \mathbf{t})$ without V. Let c_a and c_v indicate whether the first and recovered answers are correct, respectively. We assign

$$R_{\text{suf}}(c_a, c_v) = \min\big(1,\ 0.5(c_a + c_v)\big). \tag{8}$$

Full credit requires a correct first answer and a recoverable trace; half-credit supplies a signal before both events become reliable. The evolving verifier stays aligned with the policy's reasoning language.

4.3 Fidelity Anchoring

Each backbone uses its own frozen RL initialization π_0 to score $(V, Q, \mathbf{t}_i)$ as $s_i = \widehat{\mathcal{F}}_{\pi_0}(\mathbf{t}_i, V, Q) \in \{1, \ldots, 5\}$: from absent or vague visual content to several concrete, image-checkable details (Appendix B). This image-conditioned preference complements recovery without serving as a claim-level truth label. Within each group of $N = 8$ rollouts, let $m = s_{(5)}$ be the upper median and M the

maximum:

$$R_{\text{fid}}(i) = \begin{cases} \eta \dfrac{s_i - m}{M - m}, & a_i = y^* \wedge s_i > m, \\ 0, & \text{otherwise}, \end{cases} \qquad \eta = 0.1. \tag{9}$$

If $M = m$, every bonus is zero. The bonus uses relative distances above the median and is invariant to shared offsets and positive rescaling. It directs additional credit toward correct chains with stronger visual evidence under the frozen observer's preference.

4.4 Format and Length Rewards

We require `<think>...</think><answer>...</answer>`, assigning $R_{\text{format}} = 0$ for valid format and -1 otherwise. A piecewise linear penalty controls trace length:

$$R_{\text{length}} = \begin{cases} -\alpha(L_{\text{low}} - |\mathbf{t}|), & |\mathbf{t}| < L_{\text{low}}, \\ 0, & L_{\text{low}} \le |\mathbf{t}| \le L_{\text{high}}, \\ -\alpha(|\mathbf{t}| - L_{\text{high}}), & |\mathbf{t}| > L_{\text{high}}. \end{cases} \tag{10}$$

The slope α penalizes deviations without rewarding verbosity inside the interval. RL starts directly from the instruction-tuned checkpoint; Algorithm 1 summarizes training.

5 Experiments

5.1 Experimental Setup

Data and training. Bifocal-24K unifies roughly 24K examples from 19 science, visual-understanding, and multimodal-math sources. We retain compatible multiple-choice items and standardize their answer spaces without augmentation (Figure 3(a); Appendix C). We train Qwen2.5-VL-3B/7B (Bai et al., 2025b) and Qwen3-VL-4B/8B (Bai et al., 2025a) for 367 steps on 4–8 A800s, using learning rate 10^{-6}, batch size 64, eight rollouts, and KL coefficient 0.001. Both second passes use greedy decoding; Appendix B gives further settings.

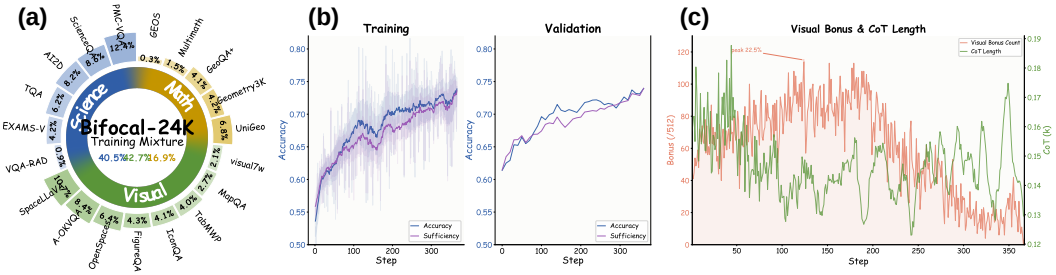


Figure 3: Training data and optimization. (a) Bifocal-24K sources. (b) Training and validation accuracy and sufficiency. (c) Visual-bonus activity and CoT length.

Evaluation. Table 1 covers 15 benchmarks and comparable-scale outcome, self-verification, and perception-enhanced baselines. Base checkpoints use default VLMEvalKit (Duan et al., 2024) prompts; reasoning baselines receive a concise common reasoning instruction. Appendix C lists benchmarks, baseline groups, and answer-checking protocols.

Training dynamics. Validation accuracy and sufficiency improve together (Figure 3(b)). Visual-bonus activation rises from 8.6% to 22.5% at step 124, then falls below 1% while accuracy remains high and CoT length stays bounded. An early increase in correct rollouts gives the observer more successful chains to distinguish. The later decline is consistent with those chains increasingly sharing its preferred evidence style, leaving fewer above-median differences. This shift emerges under a fixed two-view protocol; score ties and the correctness gate also shape bonus activation (Appendix B.5).

Table 1: Main results across 15 benchmarks. HallusionB reports aAcc; Avg is the arithmetic mean of the 15 displayed scores. Within each parameter group, including the base model, the highest value is **bold** and the second-highest distinct value is underlined; ties share the same formatting.

	Science/Math					General				Extraction				Hallucination		
Model	MMMU	MMMU-Pro	MathVerse	MathVision	ScienceQA	MMStar	MMBench	RealWorldQA	BLINK	AI2D	ChartQA	InfoVQA	OCRBench	HallusionB	POPE	Avg
3B: Qwen2.5-VL																
Qwen2.5-VL	41.1	31.7	28.6	19.2	80.2	55.9	79.6	65.5	48.9	81.8	83.8	75.9	82.3	60.7	83.8	61.3
Visionary-R1	48.2	33.8	35.9	23.1	84.7	58.6	80.4	65.8	48.0	82.1	84.6	75.8	**83.4**	65.5	80.7	63.4
Vision-SR1	50.0	**36.2**	33.9	23.3	83.4	58.2	78.9	65.1	48.0	79.5	83.4	74.8	80.2	66.0	83.9	63.0
PeBR-R1	45.2	35.2	29.6	21.8	81.9	58.9	79.8	65.8	**50.0**	81.9	**84.9**	75.8	82.0	65.0	83.7	62.8
PRCO	49.8	35.5	**39.5**	**27.5**	**85.8**	**60.9**	**81.3**	66.8	49.0	82.1	80.9	75.6	82.7	63.3	80.7	64.1
Perceval	46.4	33.9	33.8	21.6	83.5	59.0	80.2	66.3	49.1	82.3	81.2	75.5	82.9	63.5	**86.8**	63.1
Bifocal	**51.1**	35.8	37.1	24.6	85.1	60.1	80.8	**66.9**	49.6	**82.6**	**84.9**	**76.1**	82.8	**69.4**	84.9	**64.8**
4B: Qwen3-VL																
Qwen3-VL	52.8	35.9	32.3	22.4	91.6	62.9	84.5	72.7	**63.7**	84.0	**81.1**	**78.4**	**84.5**	70.6	**88.7**	67.1
Bifocal	**53.8**	**37.2**	**46.4**	**31.1**	**93.0**	**63.5**	**85.2**	**73.1**	63.5	**86.0**	80.5	78.3	83.5	**71.0**	85.5	**68.8**
7B: Qwen2.5-VL																
Qwen2.5-VL	51.6	38.4	34.4	22.0	86.2	64.3	84.5	71.4	57.4	85.3	**85.1**	82.6	87.8	71.4	85.9	67.2
Vision-R1	52.4	39.5	38.4	27.8	88.6	66.5	85.0	70.3	57.5	85.2	84.5	82.3	86.3	64.7	88.7	67.8
MM-Eureka	52.2	40.4	41.5	27.4	86.4	66.8	84.0	68.9	55.9	84.6	79.1	81.6	85.4	69.2	85.3	67.2
VL-Rethinker	54.1	42.3	42.8	28.7	91.1	64.7	84.2	69.3	56.7	84.5	79.7	81.3	86.7	70.5	84.9	68.1
Perception-R1	53.4	42.1	41.8	28.7	91.4	66.1	84.8	69.2	59.3	85.4	80.0	83.6	86.5	70.1	86.1	68.6
Vision-SR1	55.7	42.0	38.8	27.4	89.1	64.7	82.0	69.0	55.5	83.0	78.4	80.2	85.2	69.8	84.9	67.0
PeBR-R1	52.2	41.5	39.0	28.0	88.9	64.3	82.9	69.7	57.1	85.1	79.7	82.8	86.3	70.6	87.6	67.7
VPPO	55.7	42.3	44.0	30.1	88.9	66.1	84.9	69.5	**60.0**	84.5	78.5	83.6	88.5	**71.6**	86.1	69.0
PEARL	53.0	42.4	43.7	29.3	91.4	66.6	84.8	71.0	58.1	85.2	83.1	83.6	87.9	69.5	**89.4**	69.3
PRCO	53.4	43.5	**44.9**	29.7	93.6	**68.4**	86.0	69.9	59.3	85.4	81.5	**83.8**	**89.5**	70.9	86.1	69.7
Bifocal	**56.0**	**44.0**	44.6	**34.3**	**94.3**	67.2	**86.9**	**71.8**	59.4	**87.0**	83.4	83.1	87.9	71.2	87.8	**70.6**
8B: Qwen3-VL																
Qwen3-VL	53.9	40.5	49.5	31.1	93.2	64.7	85.0	72.7	67.8	85.7	82.4	82.6	87.0	73.2	87.3	70.4
VPPO	58.4	**49.3**	52.4	33.6	96.5	**72.9**	87.1	74.1	66.6	86.0	79.3	**86.2**	**88.8**	**73.8**	**89.1**	72.9
PEARL	58.0	46.2	50.5	32.1	**96.8**	72.1	86.3	73.6	67.0	84.6	82.8	83.9	87.9	71.4	87.5	72.0
PRCO	**59.3**	47.6	51.9	33.6	96.4	72.5	**88.1**	75.0	68.4	86.8	84.7	83.1	88.3	72.6	**89.1**	73.2
Bifocal	**59.3**	45.2	**54.7**	**35.2**	96.1	71.3	86.3	**77.5**	**68.7**	**88.7**	**85.2**	83.5	87.0	73.5	89.0	**73.4**

5.2 Main Results

Table 1 shows stronger reasoning across model scales while retaining competitive visual extraction performance on tasks where some reasoning baselines regress, consistent with jointly assessing answer sufficiency and visual fidelity.

Qualitative cases. Figure 4 contrasts saved 3B responses in which visual evidence, reasoning, and the final answer diverge. One baseline reaches the correct answer while citing a false food-web relation; in another case, a correctly described chart value changes in the reasoning that produces the answer. These examples show why the answer-producing chain, rather than a separate description or answer alone, is the object of both checks.

5.3 Ablation Study

To separate the contributions of answer-chain recovery and visual anchoring, Table 2 compares their configurations on five benchmarks spanning reasoning, diagrams, text extraction, and OCR.

Algorithm 1 Bifocal training

Require: Frozen π_0, policy π_θ, data $\mathcal{D}$; $N = 8$, $\eta = 0.1$, λ

for each training step **do**
  Sample a batch from $\mathcal{D}$
  for each query (x, y^*) **do**
    Sample $\{(\mathbf{t}_i, a_i)\}_{i=1}^{N}$ from $\pi_\theta(\cdot \mid x)$
    Recover $\tilde{a}_i$ from $(Q, \mathcal{C}, \mathbf{t}_i)$ with π_θ, without V
    Rate $s_i \leftarrow \pi_0(V, Q, \mathbf{t}_i)$
    Compute raw R_i using Eqs. 7–10
    $\hat{A}_i \leftarrow (R_i - \mu_R)/(\sigma_R + \delta)$
  end for
  Update π_θ by Eq. 6; keep π_0 frozen
end for
return π_θ

Table 2: 3B ablation. Full: Bifocal; Co-F: co-evolving fidelity; No-V: no visual reward; Fr-S: frozen sufficiency; GRPO: Pure GRPO. R-Avg averages MMMU, MMStar, AI2D; E-Avg averages TextVQA (Singh et al., 2019), OCRBench. C/F denote co-evolving/frozen verifiers.

	Full	Co-F	No-V	Fr-S	GRPO
Suf.	C	C	C	F	—
Visual	π_0	C	—	—	—
MMMU	**51.1**	48.2	50.0	48.7	46.8
R-Avg	**64.6**	62.4	63.5	62.5	60.4
E-Avg	**77.3**	75.2	74.6	74.6	75.3

The sufficiency-only model improves R-Avg by 3.1 points over Pure GRPO, but its E-Avg is 0.7 points lower. Adding the frozen visual branch raises E-Avg by 2.7 points and R-Avg by another 1.1. Replacing the frozen visual observer with a co-evolving one lowers R-Avg by 2.2 and E-Avg by 2.1. A co-evolving sufficiency verifier also outperforms a frozen verifier on R-Avg (63.5 vs. 62.5). These comparisons separate the reasoning benefit of direct chain recovery from the visual branch's larger contribution on extraction tasks.

5.4 Paired Tests of the Two Views

We test what each view measures through paired 3B interventions (Table 3; Appendix D). On 60 controlled OCR/table images, corrupting one fact lowers π_0's rating in 59 pairs; replacing only the image lowers it in all 60. The observer therefore uses the image to distinguish visual evidence, beyond rewarding concrete wording alone. In 38 longer AI2D/MMStar chains, a single numeric substitution produces 33 ties, showing the score operates as a coarse preference over whole chains.

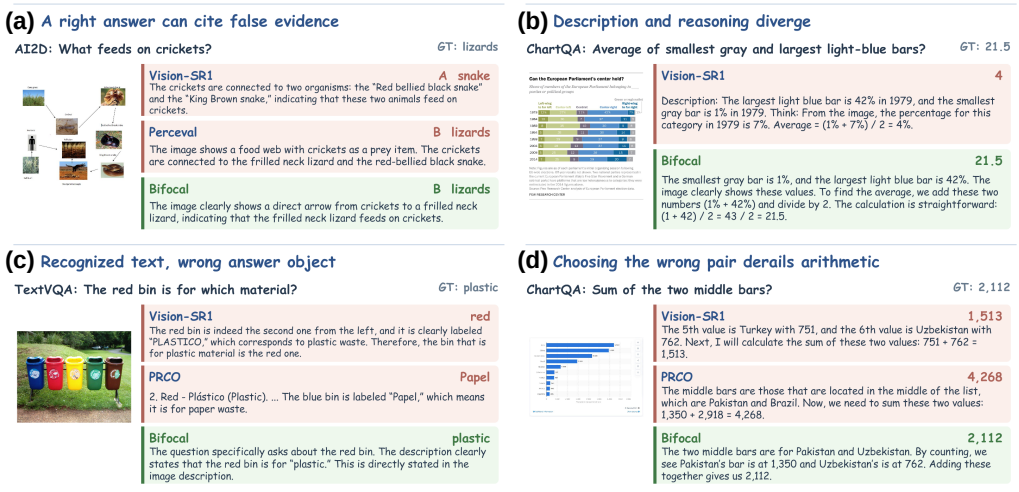


Figure 4: Answer-chain failures in saved 3B outputs. (a) A correct answer can cite a false visual relation. (b) A correct description can diverge from the subsequent reasoning. (c) Recognized text need not reach the final answer. (d) Selecting the wrong bars changes the computed result. Excerpts are lightly formatted; ellipses mark omissions.

Table 3: Paired 3B diagnostics. (a) Original rating higher than, tied with, or lower than the perturbed rating; Ctrl.: controlled images. (b) Recovery from original, empty, or mismatched chains; correct/wrong refers to the first answer. Question and options remain fixed.

(a) Frozen visual ratings

Intervention	n	$>$	$=$	$<$
Ctrl.: fact	60	59	1	0
Ctrl.: image	60	60	0	0
CoT: fact	38	3	33	2
CoT: image	38	13	24	1

(b) Answer recovery

Subset	n	Orig.	Empty	Mis.
AI2D	40	21	15	13
MMStar	40	18	5	6
MMMU	40	22	15	15
All	120	**61**	35	34
Correct	60	**55**	25	22
Wrong	60	6	10	12

Table 4: VMCBench-DEV (Zhang et al., 2025b). SR1: Vision-SR1; Bif.: Bifocal. Joint-valid requires a correct answer and $F = S = 2$. In (a), joint-valid percentages use all 200 audited items as denominator; each task has 50 items. In (b), triples list counts for scores 0/1/2 among explicitly correct chains.

(a) Task outcomes

	Accuracy (%)		Joint-valid	
Task	SR1	Bif.	SR1	Bif.
All (1,000)	75.2	**76.5**	–	–
VizWiz	76.0	**90.0**	16.0	**20.0**
MathVista	48.0	**60.0**	6.0	**7.5**
VQAv2	90.0	**94.0**	21.5	21.5
GQA	68.0	**72.0**	14.0	14.0
Four tasks	70.5	**79.0**	57.5	**63.0**

(b) Chain scores (counts)

Model	n	F	S	Joint-valid
SR1	140	11/4/125	5/16/119	115
Bif.	158	19/5/134	9/15/134	**126**

The within-question bonus thus favors overall image-conditioned evidence among successful chains, rather than detecting every isolated factual error in a long CoT.

Holding the question and options fixed, image-stripped recovery with original, empty, and mismatched chains is 61/120, 35/120, and 34/120; among initially correct answers it is 55/60, 25/60, and 22/60. The original chain therefore contributes information that the question and options alone do not provide, especially when the first answer is correct. The visual test addresses the separate issue of whether its premises are image-supported.

5.5 Visual-Evidence Diagnostics

The paired tests characterize the reward signals; we next inspect image use during answering. Let $\ell_m(V; a^*)$ denote model m's mean log-probability of the gold answer. Blurring cell c gives $\widetilde{V}_c$. For an answer-relevant region R, with $[z]_+ = \max(z, 0)$, we measure

$$s_{m,c} = \ell_m(V; a^*) - \ell_m(\widetilde{V}_c; a^*), \qquad S_m(R) = \sum_{c \in R} [s_{m,c}]_+, \qquad C_m(R) = \frac{S_m(R)}{\sum_c [s_{m,c}]_+}. \quad (11)$$

$S_m(R)$ measures positive sensitivity inside R; $C_m(R)$ is its image-wide share. In Figure 5(a), Bifocal has the largest MMStar $S_m(R)$ on decisive masses and speeds (2.74 vs. 0.74/2.14/0.11 for pre-RL/Vision-SR1/PRCO) and highest RealWorldQA $C_m(R)$ on the queried knife row (63.0% vs. 45.3%/55.5%/43.3%). PRCO has larger absolute sensitivity in the latter case but answers incorrectly. The contrast shows why selecting answer-relevant evidence and using it in the final computation are distinct demands (protocol and further cases in Appendix E).

Figure 5(b) measures $\Delta^{(r)}_{m,l} = \mathrm{logit}(v^{(r)}_{m,l}) - \mathrm{logit}(v^{(r)}_{\pi_0,l})$, where $v^{(r)}_{m,l}$ is layer-l image-attention mass from prompt or answer queries r. Vision-SR1 reduces early/middle answer-side mass; PRCO increases it. Bifocal remains near initialization on prompt queries and reduces middle/late answer-side mass. Thus its stronger answers do not require a global increase in image-token attention: panel (a) localizes where visual perturbations affect answer probability, while panel (b) shows how image attention is distributed across layers. Answer-side queries use teacher-forced gold tokens (Appendix E).

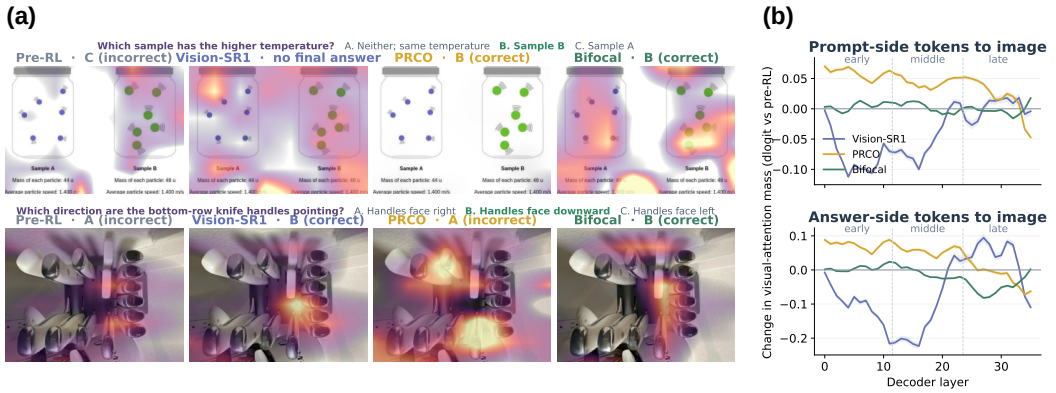


Figure 5: Visual-evidence diagnostics on Qwen2.5-VL-3B models. (a) Gold-answer sensitivity to local image blur, with a shared scale within each case. (b) Depth-wise change in visual-attention mass relative to the common pre-RL initialization over 992 inputs; bands are paired-bootstrap 95% intervals.

5.6 Additional VMCBench Tasks

On all 1,000 VMCBench-DEV items, Bifocal improves accuracy from 75.2 to 76.5. We further examine four tasks selected after comparing answer accuracy: VizWiz (Gurari et al., 2018), MathVista (Lu et al., 2024), VQAv2 (Goyal et al., 2017), and GQA (Hudson & Manning, 2019) (Table 4). Across their 200 items, correct answers rise from 141 to 158, with the largest gains on VizWiz and MathVista. One benchmark-correct Vision-SR1 output lacks an unambiguous final label, leaving 140 explicitly correct chains in panel (b).

To examine whether successful answers also satisfy both chain-level criteria, Qwen3.8-Max, accessed via API, audits both models' chains on all 200 items, scoring visual support with the image and sufficiency in a separate text-only pass. Joint validity requires a correct answer, an image-supported decisive fact ($F = 2$), and reasoning sufficient to select the answer from the question, options, and chain without the image ($S = 2$). Bifocal produces 126 such solutions versus 115 for Vision-SR1, a gain of 11 complete, visually supported answers on the same 200 questions. The gain comes from VizWiz and MathVista; VQAv2 and GQA tie despite higher answer accuracy. Thus the audit shows where improved answering also yields more chains that jointly satisfy the two desired properties. Alongside the ablations and paired interventions, it connects the training signals to the quality of the solutions produced (Appendix C.4).

6 Conclusion

Bifocal couples image-stripped recovery with a frozen pre-RL visual preference on the same answer-producing chain. Across four backbones, it strengthens reasoning while retaining visual extraction and hallucination performance. Component and paired analyses clarify their complementarity: recoverability leaves visual premises unchecked, while visual preference cannot establish that a chain carries the answer. The central design choice is therefore to assess the visual premises and their sufficiency on the very chain that receives policy-gradient credit, linking evidence selection to answer-producing reasoning.

AI USE STATEMENT

Generative AI tools assisted with methodological discussion, diagnostic experiment design, mathematical exposition and derivation checking, implementation and analysis code, and interpretation of results. They also assisted with literature discovery, citation checking, translation, manuscript editing, and figure preparation. Synthetic OCR and table cards were constructed for controlled diagnostics; the Bifocal-24K training mixture uses existing examples without synthetic augmentation. An external multimodal model supplied initial F/S judgments for the VMCBench output audit, followed by item-level review (Appendix C.4). The authors reviewed the manuscript and take responsibility for its final text, claims, code, figures, and reported results.

ETHICS STATEMENT

This work studies training and evaluation methods for vision-language models using existing datasets and benchmarks. It involves no human-subject study and no collection of private user data. The training data and automated evaluators may carry domain and demographic biases, and the resulting models can still produce inaccurate or hallucinated content. Their use in high-stakes settings requires task-specific validation and safeguards. Use and release of the data follow the licenses and terms of their original sources.

REPRODUCIBILITY STATEMENT

Sections 4 and 5.1 specify the reward definitions, training algorithm, data, and principal hyperparameters. Appendix A expands the derivations and their assumptions; Appendix B describes the two verification passes and optimization flow; Appendix C covers data processing and evaluation; Appendices D and E detail the diagnostic protocols. These sections document the scoring rubrics, controlled interventions, and statistical procedures used to interpret the results. Code, configurations, and model weights are planned for release.

REFERENCES

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, Mei Li, Kaixin Li, Zicheng Lin, Junyang Lin, Xuejing Liu, Jiawei Liu, Chenglong Liu, Yang Liu, Dayiheng Liu, Shixuan Liu, Dunjie Lu, Ruilin Luo, Chenxu Lv, Rui Men, Lingchen Meng, Xuancheng Ren, Xingzhang Ren, Sibo Song, Yuchong Sun, Jun Tang, Jianhong Tu, Jianqiang Wan, Peng Wang, Pengfei Wang, Qiuyue Wang, Yuxuan Wang, Tianbao Xie, Yiheng Xu, Haiyang Xu, Jin Xu, Zhibo Yang, Mingkun Yang, Jianxin Yang, An Yang, Bowen Yu, Fei Zhang, Hang Zhang, Xi Zhang, Bo Zheng, Humen Zhong, Jingren Zhou, Fan Zhou, Jing Zhou, Yuanzhi Zhu, and Ke Zhu. 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.

Jie Cao and Jing Xiao. An augmented benchmark dataset for geometric question answering through dual parallel text encoding. In Nicoletta Calzolari, Chu-Ren Huang, Hansaem Kim, James Pustejovsky, Leo Wanner, Key-Sun Choi, Pum-Mo Ryu, Hsin-Hsi Chen, Lucia Donatelli, Heng Ji, Sadao Kurohashi, Patrizia Paggio, Nianwen Xue, Seokhwan Kim, Younggyun Hahm, Zhong He, Tony Kyungil Lee, Enrico Santus, Francis Bond, and Seung-Hoon Na (eds.), *Proceedings of the 29th International Conference on Computational Linguistics*, pp. 1511–1520, Gyeongju, Republic of Korea, October 2022. International Committee on Computational Linguistics. URL https://aclanthology.org/2022.coling-1.130/.

Shuaichen Chang, David Palzer, Jialin Li, Eric Fosler-Lussier, and Ningchuan Xiao. Mapqa: A dataset for question answering on choropleth maps, 2022. URL https://arxiv.org/abs/2211.08545.

Hardy Chen, Haoqin Tu, Fali Wang, Hui Liu, Xianfeng Tang, Xinya Du, Yuyin Zhou, and Cihang Xie. Sft or rl? an early investigation into training r1-like reasoning large vision-language models, 2025a. URL https://arxiv.org/abs/2504.11468.

Jiaqi Chen, Tong Li, Jinghui Qin, Pan Lu, Liang Lin, Chongyu Chen, and Xiaodan Liang. UniGeo: Unifying geometry logical reasoning via reformulating mathematical expression. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang (eds.), *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 3313–3323, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.emnlp-main.218. URL https://aclanthology.org/2022.emnlp-main.218/.

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 A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 27056–27087. Curran Associates, Inc., 2024. doi: 10.52202/079017-0850. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/2f8ee6a3d766b426d2618e555b5aeb39-Paper-Conference.pdf.

Yan Chen, Long Li, Teng Xi, Long Zeng, and Jingdong Wang. Perception before reasoning: Two-stage reinforcement learning for visual reasoning in vision-language models, 2025b. URL https://arxiv.org/abs/2509.13031.

Rocktim Das, Simeon Hristov, Haonan Li, Dimitar Dimitrov, Ivan Koychev, and Preslav Nakov. EXAMS-V: A multi-discipline multilingual multimodal exam benchmark for evaluating vision language models. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 7768–7791, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.420. URL https://aclanthology.org/2024.acl-long.420/.

Haodong Duan, Junming Yang, Yuxuan Qiao, Xinyu Fang, Lin Chen, Yuan Liu, Xiaoyi Dong, Yuhang Zang, Pan Zhang, Jiaqi Wang, Dahua Lin, and Kai Chen. Vlmevalkit: An open-source toolkit for evaluating large multi-modality models. In *Proceedings of the 32nd ACM International Conference on Multimedia*, MM '24, pp. 11198–11201, New York, NY, USA, 2024. Association for Computing Machinery. ISBN 9798400706868. doi: 10.1145/3664647.3685520. URL https://doi.org/10.1145/3664647.3685520.

Xingyu Fu, Yushi Hu, Bangzheng Li, Yu Feng, Haoyu Wang, Xudong Lin, Dan Roth, Noah A. Smith, Wei-Chiu Ma, and Ranjay Krishna. Blink: Multimodal large language models can see but not perceive, 2024. URL https://arxiv.org/abs/2404.12390.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

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

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, Aixin Liu, Bing Xue, Bingxuan Wang, Bochao Wu, Bei Feng, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chong Ruan, Damai Dai, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, Fuli Luo, Guangbo Hao, Guanting Chen, Guowei Li, H. Zhang, Hanwei Xu, Honghui Ding, Huazuo Gao, Hui Qu, Hui Li, Jianzhong Guo, Jiashi Li, Jingchang Chen, Jingyang Yuan, Jinhao Tu, Junjie Qiu, Junlong Li, J. L. Cai, Jiaqi Ni, Jian Liang, Jin Chen, Kai Dong, Kai Hu, Kaichao You, Kaige Gao, Kang Guan, Kexin Huang, Kuai Yu, Lean Wang, Lecong Zhang, Liang Zhao, Litong Wang, Liyue Zhang, Lei Xu, Leyi Xia, Mingchuan Zhang, Minghua Zhang, Minghui Tang, Mingxu Zhou, Meng Li, Miaojun

Wang, Mingming Li, Ning Tian, Panpan Huang, Peng Zhang, Qiancheng Wang, Qinyu Chen, Qiushi Du, Ruiqi Ge, Ruisong Zhang, Ruizhe Pan, Runji Wang, R. J. Chen, R. L. Jin, Ruyi Chen, Shanghao Lu, Shangyan Zhou, Shanhuang Chen, Shengfeng Ye, Shiyu Wang, Shuiping Yu, Shunfeng Zhou, Shuting Pan, S. S. Li, Shuang Zhou, Shaoqing Wu, Tao Yun, Tian Pei, Tianyu Sun, T. Wang, Wangding Zeng, Wen Liu, Wenfeng Liang, Wenjun Gao, Wenqin Yu, Wentao Zhang, W. L. Xiao, Wei An, Xiaodong Liu, Xiaohan Wang, Xiaokang Chen, Xiaotao Nie, Xin Cheng, Xin Liu, Xin Xie, Xingchao Liu, Xinyu Yang, Xinyuan Li, Xuecheng Su, Xuheng Lin, X. Q. Li, Xiangyue Jin, Xiaojin Shen, Xiaosha Chen, Xiaowen Sun, Xiaoxiang Wang, Xinnan Song, Xinyi Zhou, Xianzu Wang, Xinxia Shan, Y. K. Li, Y. Q. Wang, Y. X. Wei, Yang Zhang, Yanhong Xu, Yao Li, Yao Zhao, Yaofeng Sun, Yaohui Wang, Yi Yu, Yichao Zhang, Yifan Shi, Yiliang Xiong, Ying He, Yishi Piao, Yisong Wang, Yixuan Tan, Yiyang Ma, Yiyuan Liu, Yongqiang Guo, Yuan Ou, Yuduan Wang, Yue Gong, Yuheng Zou, Yujia He, Yunfan Xiong, Yuxiang Luo, Yuxiang You, Yuxuan Liu, Yuyang Zhou, Y. X. Zhu, Yanping Huang, Yaohui Li, Yi Zheng, Yuchen Zhu, Yunxian Ma, Ying Tang, Yukun Zha, Yuting Yan, Z. Z. Ren, Zehui Ren, Zhangli Sha, Zhe Fu, Zhean Xu, Zhenda Xie, Zhengyan Zhang, Zhewen Hao, Zhicheng Ma, Zhigang Yan, Zhiyu Wu, Zihui Gu, Zijia Zhu, Zijun Liu, Zilin Li, Ziwei Xie, Ziyang Song, Zizheng Pan, Zhen Huang, Zhipeng Xu, Zhongyu Zhang, and Zhen Zhang. Deepseek-r1 incentivizes reasoning in llms through reinforcement learning. *Nature*, 645 (8081):633–638, September 2025. ISSN 1476-4687. doi: 10.1038/s41586-025-09422-z. URL http://dx.doi.org/10.1038/s41586-025-09422-z.

Danna Gurari, Qing Li, Abigale J. Stangl, Anhong Guo, Chi Lin, Kristen Grauman, Jiebo Luo, and Jeffrey P. Bigham. Vizwiz grand challenge: Answering visual questions from blind people. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

Minbin Huang, Runhui Huang, Chuanyang Zheng, Jingyao Li, Guoxuan Chen, Han Shi, and Hong Cheng. Answer-consistent chain-of-thought reinforcement learning for multi-modal large langauge models, 2025. URL https://arxiv.org/abs/2510.10104.

Siyuan Huang, Xiaoye Qu, Yafu Li, Yun Luo, Zefeng He, Daizong Liu, and Yu Cheng. Spotlight on token perception for multimodal reinforcement learning. In C. Vondrick, B. Hariharan, C. Raffel, L. Pinto, D. Yang, and A. Faust (eds.), *International Conference on Learning Representations*, volume 2026, pp. 103291–103325, 2026a. URL https://proceedings.iclr.cc/paper_files/paper/2026/file/a8e3197f627b9c88f86c0d3eb1ade5d7-Paper-Conference.pdf.

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 C. Vondrick, B. Hariharan, C. Raffel, L. Pinto, D. Yang, and A. Faust (eds.), *International Conference on Learning Representations*, volume 2026, pp. 63794–63812, 2026b. URL https://proceedings.iclr.cc/paper_files/paper/2026/file/680b2a8135b9c71278a09cafb605869e-Paper-Conference.pdf.

Drew A. Hudson and Christopher D. Manning. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

Samira Ebrahimi Kahou, Vincent Michalski, Adam Atkinson, Akos Kadar, Adam Trischler, and Yoshua Bengio. Figureqa: An annotated figure dataset for visual reasoning, 2018. URL https://arxiv.org/abs/1710.07300.

Aniruddha Kembhavi, Mike Salvato, Eric Kolve, Minjoon Seo, Hannaneh Hajishirzi, and Ali Farhadi. A diagram is worth a dozen images, 2016. URL https://arxiv.org/abs/1603.07396.

Aniruddha Kembhavi, Minjoon Seo, Dustin Schwenk, Jonghyun Choi, Ali Farhadi, and Hannaneh Hajishirzi. Are you smarter than a sixth grader? textbook question answering for multimodal machine comprehension. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

Jason J. Lau, Soumya Gayen, Asma Ben Abacha, and Dina Demner-Fushman. A dataset of clinically generated visual questions and answers about radiology images. *Scientific Data*, 5(1):180251, Nov 2018. ISSN 2052-4463. doi: 10.1038/sdata.2018.251. URL https://doi.org/10.1038/sdata.2018.251.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 292–305, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.20. URL https://aclanthology.org/2023.emnlp-main.20/.

Zongxia Li, Wenhao Yu, Chengsong Huang, Zhenwen Liang, Rui Liu, Fuxiao Liu, Jingxi Chen, Dian Yu, Jordan Boyd-Graber, Haitao Mi, and Dong Yu. Vision-sr1: Self-rewarding vision-language model via reasoning decomposition and multi-reward policy optimization. In C. Vondrick, B. Hariharan, C. Raffel, L. Pinto, D. Yang, and A. Faust (eds.), *International Conference on Learning Representations*, volume 2026, pp. 81753–81772, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/file/8442d394d8266db6fdf92583a4783e8c-Paper-Conference.pdf.

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?, 2024a. URL https://arxiv.org/abs/2307.06281.

Yuliang Liu, Zhang Li, Mingxin Huang, Biao Yang, Wenwen Yu, Chunyuan Li, Xu-Cheng Yin, Cheng-Lin Liu, Lianwen Jin, and Xiang Bai. Ocrbench: on the hidden mystery of ocr in large multimodal models. *Science China Information Sciences*, 67(12), December 2024b. ISSN 1869-1919. doi: 10.1007/s11432-024-4235-6. URL http://dx.doi.org/10.1007/s11432-024-4235-6.

Ziyu Liu, Zeyi Sun, Yuhang Zang, Xiaoyi Dong, Yuhang Cao, Haodong Duan, Dahua Lin, and Jiaqi Wang. Visual-rft: Visual reinforcement fine-tuning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 2034–2044, October 2025.

Pan Lu, Ran Gong, Shibiao Jiang, Liang Qiu, Siyuan Huang, Xiaodan Liang, and Song-Chun Zhu. Inter-GPS: Interpretable geometry problem solving with formal language and symbolic reasoning. In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli (eds.), *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 6774–6786, Online, August 2021a. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.528. URL https://aclanthology.org/2021.acl-long.528/.

Pan Lu, Liang Qiu, Jiaqi Chen, Tanglin Xia, Yizhou Zhao, Wei Zhang, Zhou Yu, Xiaodan Liang, and Song-Chun Zhu. Iconqa: A new benchmark for abstract diagram understanding and visual language reasoning. In J. Vanschoren and S. Yeung (eds.), *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, volume 1, 2021b. URL https://datasets-benchmarks-proceedings.neurips.cc/paper_files/paper/2021/file/d3d9446802a44259755d38e6d163e820-Paper-round2.pdf.

Pan Lu, Swaroop Mishra, Tanglin Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. Learn to explain: Multimodal reasoning via thought chains for science question answering. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 2507–2521. Curran Associates, Inc., 2022. doi: 10.52202/068431-0182. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/11332b6b6cf4485b84afadb1352d3a9a-Paper-Conference.pdf.

Pan Lu, Liang Qiu, Kai-Wei Chang, Ying Nian Wu, Song-Chun Zhu, Tanmay Rajpurohit, Peter Clark, and Ashwin Kalyan. Dynamic prompt learning via policy gradient for semi-structured mathematical reasoning, 2023. URL https://arxiv.org/abs/2209.14610.

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 B. Kim, Y. Yue, S. Chaudhuri, K. Fragkiadaki, M. Khan, and Y. Sun (eds.), *International Conference on Learning Representations*, volume 2024, pp. 23439–23554, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/file/663bce02a0050c4a11f1eb8a7f1429d3-Paper-Conference.pdf.

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 Smaranda Muresan, Preslav Nakov, and Aline Villavicencio (eds.), *Findings of the Association for Computational Linguistics: ACL 2022*, pp. 2263–2279, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.findings-acl.177. URL https://aclanthology.org/2022.findings-acl.177/.

Minesh Mathew, Viraj Bagal, Rubèn Tito, Dimosthenis Karatzas, Ernest Valveny, and C.V. Jawahar. Infographicvqa. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 1697–1706, January 2022.

Fanqing Meng, Lingxiao Du, Zongkai Liu, Zhixiang Zhou, Quanfeng Lu, Tiancheng Han, Daocheng Fu, Kaipeng Zhang, Ping Luo, Yu Qiao, Jiaheng Zhang, Michael Qizhe Shieh, Qiaosheng Zhang, and Wenqi Shao. MM-eureka: Toward stable multimodal reasoning via rule-based reinforcement learning with policy drift control. *Transactions on Machine Learning Research*, 2026. ISSN 2835-8856. URL https://openreview.net/forum?id=8y1ch6y24H.

Ziqi Miao, Haonan Jia, Lijun Li, Chen Qian, Yuan Xiong, Wenting Yan, and Jing Shao. Seeing with you: Perception-reasoning coevolution for multimodal reasoning, 2026. URL https://arxiv.org/abs/2603.28618.

Yingqian Min, Kun Zhou, Yifan Li, Yuhuan Wu, Han Peng, Yifan Du, Wayne Xin Zhao, Min Yang, and Ji-Rong Wen. Improving vision-language models with perception-centric process reward models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 33099–33109, June 2026.

Shuai Peng, Di Fu, Liangcai Gao, Xiuqin Zhong, Hongguang Fu, and Zhi Tang. Multimath: Bridging visual and mathematical reasoning for large language models, 2024. URL https://arxiv.org/abs/2409.00147.

remyxai. Vqasynth, 2024. URL https://github.com/remyxai/VQASynth/tree/main. GitHub repository.

Dustin Schwenk, Apoorv Khandelwal, Christopher Clark, Kenneth Marino, and Roozbeh Mottaghi. A-okvqa: A benchmark for visual question answering using world knowledge, 2022. URL https://arxiv.org/abs/2206.01718.

Minjoon Seo, Hannaneh Hajishirzi, Ali Farhadi, Oren Etzioni, and Clint Malcolm. Solving geometry problems: Combining text and diagram interpretation. In Lluís Màrquez, Chris Callison-Burch, and Jian Su (eds.), *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pp. 1466–1476, Lisbon, Portugal, September 2015. Association for Computational Linguistics. doi: 10.18653/v1/D15-1171. URL https://aclanthology.org/D15-1171/.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. Deepseekmath: Pushing the limits of mathematical reasoning in open language models, 2024. URL https://arxiv.org/abs/2402.03300.

Haozhan Shen, Peng Liu, Jingcheng Li, Chunxin Fang, Yibo Ma, Jiajia Liao, Qiaoli Shen, Zilun Zhang, Kangjia Zhao, Qianqian Zhang, Ruochen Xu, and Tiancheng Zhao. Vlm-r1: A stable and generalizable r1-style large vision-language model, 2025. URL https://arxiv.org/abs/2504.07615.

Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. Towards vqa models that can read. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

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. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen (eds.), *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pp. 30865–30891. Curran Associates, Inc., 2025. doi: 10.52202/085713-1037. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/2c84844a559e4f962752570bff456ae4-Paper-Conference.pdf.

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 A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 95095–95169. Curran Associates, Inc., 2024. doi: 10.52202/079017-3014. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/ad0edc7d5fa1a783f063646968b7315b-Paper-Datasets_and_Benchmarks_Track.pdf.

Jiaer Xia, Yuhang Zang, Peng Gao, Sharon Li, and Kaiyang Zhou. Visionary-r1: Mitigating shortcuts in visual reasoning with reinforcement learning. *Transactions on Machine Learning Research*, 2026. ISSN 2835-8856. URL https://openreview.net/forum?id=JWkZXBgh5a.

Tong Xiao, Xin Xu, Zhenya Huang, Hongyu Gao, Quan Liu, Qi Liu, and Enhong Chen. Perception-r1: Advancing multimodal reasoning capabilities of mllms via visual perception reward. In C. Vondrick, B. Hariharan, C. Raffel, L. Pinto, D. Yang, and A. Faust (eds.), *International Conference on Learning Representations*, volume 2026, pp. 26868–26898, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/file/2cf153951b5e9b39564fc4a0ef6adc1a-Paper-Conference.pdf.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, Cong Wei, Botao Yu, Ruibin Yuan, Renliang Sun, Ming Yin, Boyuan Zheng, Zhenzhu Yang, Yibo Liu, Wenhao Huang, Huan Sun, Yu Su, and Wenhu Chen. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9556–9567, June 2024.

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 Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 15134–15186, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.736. URL https://aclanthology.org/2025.acl-long.736/.

Chi Zhang, Haibo Qiu, Qiming Zhang, Yufei Xu, Zhixiong Zeng, Siqi Yang, Peng Shi, Lin Ma, and Jing Zhang. Perceptual-evidence anchored reinforced learning for multimodal reasoning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 41111–41120, June 2026.

Jingyi Zhang, Jiaxing Huang, Huanjin Yao, Shunyu Liu, Xikun Zhang, Shijian Lu, and Dacheng Tao. R1-vl: Learning to reason with multimodal large language models via step-wise group relative policy optimization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 1859–1869, October 2025a.

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?, 2024a. URL https://arxiv.org/abs/2403.14624.

Xiaoman Zhang, Chaoyi Wu, Ziheng Zhao, Weixiong Lin, Ya Zhang, Yanfeng Wang, and Weidi Xie. Pmc-vqa: Visual instruction tuning for medical visual question answering, 2024b. URL https://arxiv.org/abs/2305.10415.

Yuhui Zhang, Yuchang Su, Yiming Liu, Xiaohan Wang, James Burgess, Elaine Sui, Chenyu Wang, Josiah Aklilu, Alejandro Lozano, Anjiang Wei, Ludwig Schmidt, and Serena Yeung-Levy. Automated generation of challenging multiple-choice questions for vision language model evaluation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 29580–29590, June 2025b.

Yuke Zhu, Oliver Groth, Michael Bernstein, and Li Fei-Fei. Visual7w: Grounded question answering in images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

A Additional Analysis

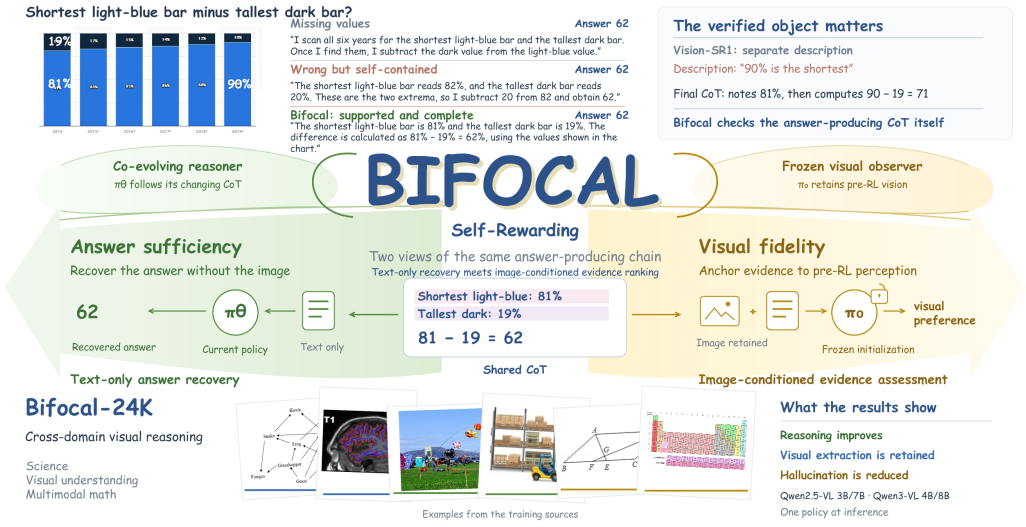


Figure 6: Bifocal Self-Rewarding assesses the answer-producing chain from two views, accompanied by Bifocal-24K and improvements across multiple backbones. The first two ChartQA traces illustrate evidence deficiencies; the Bifocal and Vision-SR1 excerpts show model outputs.

This section develops the two-view formulation in Section 3, derives its contribution to the policy update, and explains the ordering induced by the composite reward. The analysis distinguishes the target properties of a chain from the scores used to optimize and evaluate them.

A.1 Neither Property Implies the Other

Consider idealized binary properties $S(\mathbf{t})$ and $F(\mathbf{t}, V)$ corresponding to the two targets in Eq. 2. A chain can state a false visual premise yet still determine the correct option, giving $S = 1, F = 0$. Another can list only accurate but answer-irrelevant observations, giving $S = 0, F = 1$. These counterexamples establish that neither property implies the other. The desired set is their intersection:

$$\{\mathbf{t} : S(\mathbf{t}) = 1, F(\mathbf{t}, V) = 1\} = \{\mathbf{t} : S(\mathbf{t}) = 1\} \cap \{\mathbf{t} : F(\mathbf{t}, V) = 1\}. \tag{12}$$

As long as both counterexample types are possible, either single-property set strictly contains this intersection. Stronger verification of one property alone therefore cannot guarantee the other.

A.2 Complementary Image Access

The main text suppresses option-label bookkeeping: $\mathcal{C}$ denotes the options supplied to each pass. In the implementation, y^* is the original gold label, $\widetilde{\mathcal{C}}$ is the permuted option set, and $\tilde{y}^*$ is its remapped gold label. Thus $c_a = \mathbf{1}[a = y^*]$ and $c_v = \mathbf{1}[\tilde{a} = \tilde{y}^*]$ compare each answer against its own label mapping.

Suppose a verifier retains the image and predicts $\tilde{a} = g(V, Q, \mathcal{C}, \mathbf{t})$. It may recover y^* directly from $(V, Q, \mathcal{C})$ even when its output is invariant to $\mathbf{t}$, so success cannot identify what the chain contributes. The image-stripped pass instead estimates

$$\Pr\big(g_{\theta_k}(Q, \widetilde{\mathcal{C}}, \mathbf{t}) = \tilde{y}^*\big). \tag{13}$$

For fixed $(Q, \widetilde{\mathcal{C}}, \mathbf{t})$, changing V cannot alter this second-pass distribution: the direct image-to-answer path has been removed. This does not make the chain the only information source, because the question and options remain available. Option permutation breaks a fixed correspondence with the first-pass letter; the original/empty/mismatched-chain comparison in Table 3 tests whether the original chain adds information beyond those shared inputs.

The reverse information restriction applies to fidelity. For a text-only scorer $h(Q, \mathbf{t})$, holding its inputs fixed while changing the unseen image gives

$$R_h(\mathbf{t}; V) \equiv h(Q, \mathbf{t}) = R_h(\mathbf{t}; V') \quad \text{for all } V, V'. \tag{14}$$

The support of a claim in $\mathbf{t}$ may differ between V and V', yet the text-only score cannot. Supplying V makes that distinction possible, not automatic: the frozen observer's rating remains a proxy. Thus the two interventions remove a direct visual bypass for recovery and restore the image for evidence assessment, respectively.

A.3 Verification and Policy-Gradient Credit

Let $\mathbf{d}$ be an auxiliary description and $\mathbf{t}$ the chain that directly precedes the answer. A certificate on $\mathbf{d}$ need not transfer to $\mathbf{t}$: the downstream mapping $\mathbf{t} = f(V, Q, \mathbf{d})$ may omit a correct observation in $\mathbf{d}$, or introduce an unsupported one. Hence

$$\mathcal{S}(\mathbf{d}) \text{ high} \not\Rightarrow \mathcal{S}(\mathbf{t}) \text{ high}. \tag{15}$$

Bifocal evaluates both rewards on the sampled $\mathbf{t}$ that belongs to the updated response. The visual premises and their use in answering are therefore assessed on the trajectory that receives policy-gradient credit. The token-level derivation below makes this correspondence explicit.

A.4 Token-Level Policy Update and Detached Verification

Equation 6 uses compact trajectory-level notation. The implemented surrogate uses tokenwise ratios, as in the original GRPO formulation (Shao et al., 2024). Write the sampled response as $\mathbf{z}_i = (z_{i,1}, \ldots, z_{i,L_i}) = (\mathbf{t}_i, a_i)$ and let

$$\rho_{i,j} = \frac{\pi_\theta(z_{i,j} \mid x, z_{i,<j})}{\pi_{\theta_k}(z_{i,j} \mid x, z_{i,<j})}, \qquad \bar{\rho}_{i,j} = \mathrm{clip}(\rho_{i,j}, 1-\epsilon, 1+\epsilon). \tag{16}$$

At update k, both the image-stripped verifier g_{θ_k} and the frozen observer h_{θ_0} produce rewards before the policy update. The resulting group-normalized $\widehat{A}_i^{\mathrm{Bifocal}}$ is detached; no gradient is propagated through either second pass. With $u_{i,j} = \min(\rho_{i,j}\widehat{A}_i, \bar{\rho}_{i,j}\widehat{A}_i)$, token-mean aggregation gives the clipped policy term

$$\mathcal{J}_{\mathrm{PG}}^{\mathrm{tok}}(\theta) = \mathbb{E}_{x,G\sim\pi_{\theta_k}}\left[\frac{\sum_i \sum_{j=1}^{L_i} \psi_{i,j}}{\sum_i L_i}\right], \qquad \psi_{i,j} = \begin{cases} u_{i,j}, & \widehat{A}_i \geq 0, \\ \max(u_{i,j}, c\widehat{A}_i), & \widehat{A}_i < 0, \end{cases} \tag{17}$$

where $c > 1$ is the dual-clipping factor for negative advantages and $\widehat{A}_i = \mathrm{sg}(\widehat{A}_i^{\mathrm{Bifocal}})$ inside $u_{i,j}$ and $\psi_{i,j}$. The optimizer additionally subtracts β times a token-mean KL estimate toward π_0. At $\theta = \theta_k$, the ratios equal one and clipping is inactive. Omitting the separate KL term then gives the local policy-gradient contribution

$$\nabla_\theta \mathcal{J}_{\mathrm{PG}}^{\mathrm{tok}}\big|_{\theta_k} = \mathbb{E}\left[\frac{1}{\sum_i L_i} \sum_i \widehat{A}_i^{\mathrm{Bifocal}} \sum_{j=1}^{L_i} \nabla_\theta \log \pi_\theta(z_{i,j} \mid x, z_{i,<j})\big|_{\theta_k}\right]. \tag{18}$$

The same trajectory-level advantage multiplies the token gradients of its answer-producing chain and final answer. Co-evolution changes g_{θ_k} between updates; it does not add a differentiable path through the second rollout within an update.

A.5 Reward Ordering Before Group Standardization

We use $\lambda = 0.7$ in Eq. 7. Because $c_a, c_v \in \{0, 1\}$, Eq. 8 simplifies exactly to $R_{\mathrm{suf}} = (c_a + c_v)/2$. Writing $b_i = R_{\mathrm{fid}}(i)$ and $u_i = R_{\mathrm{format},i} + R_{\mathrm{length},i}$, Eq. 7 therefore has the pre-standardization form

$$R_i = 0.85c_{a,i} + 0.15c_{v,i} + b_i + u_i, \qquad 0 \leq b_i \leq 0.1, \tag{19}$$

with $b_i = 0$ whenever $c_{a,i} = 0$. For two rollouts with equal auxiliary terms, a correct first answer exceeds an incorrect one by at least $0.85 - 0.15 = 0.70$, even if the latter passes the text-only check.

Among two first-pass-correct rollouts, a successful image-stripped recovery exceeds a failed one by at least $0.15 - 0.10 = 0.05$. When both correctness indicators agree, the visual bonus can determine their ordering. Group standardization preserves these signs because

$$\widehat{A}_i - \widehat{A}_j = \frac{R_i - R_j}{\sigma_{G(x)} + \delta}. \tag{20}$$

The bound on b_i is a bound on reward scale before standardization, not on the magnitude of the resulting advantage. If all other reward terms tie, the visual bonus can supply the entire within-group ordering, as intended. Different auxiliary penalties can change the pairwise margins above; the statements are conditional on equal u_i.

A.6 Calibration-Free Local Preference

For $N = 8$, write the ordered scores as $s_{(1)} \leq \cdots \leq s_{(8)}$. Equation 9 uses $m = s_{(5)}$ and $M = s_{(8)}$. Thus $M = m$ precisely when at least four scores share the group maximum; all eight need not tie. If $M > m$, at most three rollouts can score above m, and the correctness gate may reduce the number receiving a bonus further. Consequently, the logged bonus count alone cannot separate score ties from correctness-gate effects.

Let a scorer transformation be $s'_i = as_i + b$ with $a > 0$. Its median and maximum become $m' = am + b$ and $M' = aM + b$, while $s'_i > m'$ if and only if $s_i > m$. For every eligible rollout with $M > m$,

$$\frac{s'_i - m'}{M' - m'} = \frac{a(s_i - m)}{a(M - m)} = \frac{s_i - m}{M - m}. \tag{21}$$

When $M = m$, both the original and transformed groups receive zero bonus, so invariance also holds in the degenerate case. Moreover, $m < s_i \leq M$ implies $0 < R_{\mathrm{fid}}(i) \leq \eta$, and an incorrect answer receives zero. This is a local preference rule, not a guarantee that the rating orders chains by their ideal claim-level fidelity.

Table 5 distinguishes the target property, the training preference, and the output audit. The audit uses an external rubric on completed solutions, so its criterion is specified separately from the reward used during training.

Table 5: Roles of the visual quantities used in the paper.

Quantity	Role	Interpretation
$\mathcal{F}(\mathbf{t}, V)$	Target property	Support of the chain's visual claims by the image.
$s_i \in \{1, \ldots, 5\}$	Training preference	Frozen π_0 rates the evidence exposed by a chain; Eq. 9 converts within-question differences into a bonus.
$F \in \{0, 1, 2\}$	Output audit	An external judge checks decisive visual facts, alongside a separate text-only sufficiency rating S.

A.7 Asymmetric Checkpoint Roles

At update k, the trace distribution is generated by π_{θ_k}. The image-stripped verifier g_{θ_k} follows the policy's current reasoning language, while the visual observer remains the pre-RL h_{θ_0}:

$$s_i^{(k)} = h_{\theta_0}(V, Q, \mathbf{t}_i^{(k)}), \tag{22}$$

For any fixed input $(V, Q, \mathbf{t})$, the observer's parameters and scoring instruction remain unchanged across updates. Changes in its ratings during training therefore reflect changes in the sampled trajectories under a fixed evaluator. The policy begins at $\pi_{\theta_0} = \pi_0$ and is KL-regularized toward that checkpoint, so the two models share an initial visual and linguistic interface. KL controls policy drift, whereas the visual preference distinguishes successful trajectories within a question. The checkpoint ablation evaluates this choice, and the paired tests in Section 5.4 characterize the observer's image dependence and whole-chain resolution.

B Implementation Notes

B.1 Shared Trajectory Interface

For every first-pass response, the reasoning enclosed by the designated think field is treated as the answer-producing chain. Both second passes receive this same chain and return a compact structured output. They neither revise the first response nor introduce a separate description trajectory. The recovered answer determines c_v, whereas the visual rating supplies s_i; the rating is not itself the fidelity reward. Each prompt interface is specified below by its inputs, task requirements, and output fields.

First-pass rollout. The policy receives the image, question, and labeled choices. The instruction requests step-by-step reasoning followed by the selected option letter, using a single think field and a final answer field:

```
Input: image; question; labeled choices.
Task: reason through the question, then select an
option.
Output fields:
<think>reasoning </think>
<answer>option label </answer>
```

The ground-truth label is held outside the generation prompt for reward computation. Eight first-pass responses are sampled for each question. The complete think field of each response is then reused by both evaluators, rather than being replaced by a newly generated description. The two subsequent passes are greedy; this distinction separates candidate generation from reward evaluation.

B.2 Sufficiency Prompt

The current policy receives the question, permuted options, and the chain without the image. The correct option is remapped after permutation. Its task is to recover an answer, not to assign a subjective sufficiency score:

```
Answer the question from the supplied reasoning record
and the listed choices. The image is unavailable; use
the record as the source of visual information.
Select the matching choice under the labels shown
below.
TASK AND CHOICES:
{question; relabeled choices}
REASONING RECORD:
{answer-producing chain}
Provide the selected label in the answer field:
<answer>[label] </answer>
```

The parser reads the option label from the answer field and compares it with the remapped target. Option permutation changes the mapping between labels and answer content. The verifier must identify the option under this mapping; the question, option content, and any answer information already written in the chain remain available. The paired chain-removal and mismatch controls measure the chain's added contribution.

B.3 Fidelity Prompt

The frozen RL initialization receives the image, question, and chain. Its five-level rating guidance distinguishes the amount of concrete, image-checkable visual detail exposed by the chain rather than scoring answer correctness:

```
Read the reasoning alongside the image and question.
Rate how much concrete visual information the
reasoning makes available for checking against the
image.
Rating guide:
1:  visual information is essentially absent;
2:  references to the image remain generic;
3:  at least one concrete visual detail is stated;
4:  two or three concrete details are stated;
5:  three or more specific details can be checked
against the image.
IMAGE: {image}
QUESTION: {question}
REASONING TRACE: {reasoning}
Respond with a single integer in the range 1--5.
```

The parser accepts only an integer in the specified range. The references to detail counts guide the observer toward specific evidence, while the image and question provide the context for its whole-chain judgment. The upper levels overlap because the rubric expresses degrees of evidential specificity rather than a deterministic count-to-score function. During optimization, the score intervals above the group median determine the visual bonus. Section 5.4 examines how this preference responds to image and fact perturbations.

B.4 Reward and Update Flow

Answer correctness first determines whether a rollout solves the task. The image-stripped pass then measures whether the same chain retains enough answer-determining content, while the frozen image-conditioned pass ranks the visual evidence exposed by successful chains. The fidelity contribution is activated only within the eligible portion of a rollout group and, for equal auxiliary penalties, remains subordinate to task correctness. Format and length terms suppress malformed or degenerate trajectories. The resulting composite rewards are normalized within each group and passed to the standard clipped GRPO update. No separately trained reward model is introduced.

The upper median and maximum in Eq. 9 are computed over all eight ratings for a question, before applying the correctness gate. Thus incorrect rollouts receive no visual bonus but still participate in the group statistics. The resulting bonus is added to the correctness–sufficiency and auxiliary terms before reward standardization. Format and length rewards retain exactly the definitions in Section 4: valid formatting and in-range length incur zero penalty, and neither term introduces a positive reward for verbosity.

Table 6 collects the principal settings. The same pre-RL checkpoint defines the frozen visual observer and KL reference throughout training; the policy alone is updated. Batch size denotes the global number of questions, not the per-GPU batch: 64 questions with eight rollouts yield 512 first-pass trajectories. GPU count specifies the hardware allocation; it does not multiply the global batch or the number of candidates compared within each question.

Table 6: Principal Bifocal training settings.

Setting	Value
Learning rate	10^{-6}
Training steps	367
Global question batch size	64
Rollouts per question	8
KL coefficient β	0.001
Correctness–sufficiency weight λ	0.7
Maximum visual bonus η	0.1
Second-pass decoding	Greedy
Training GPUs	4 A800s (3B/4B); 8 A800s (7B/8B)

B.5 Training Settings and Bonus Activity

The frozen visual scorer is capped at 16 output tokens. Both second passes retain the same greedy-decoding protocol throughout training, allowing the logged bonus activity to be interpreted under a fixed evaluation procedure.

Figure 3(c) tracks how often the visual bonus distinguishes rollouts. Early improvements in answer correctness make more trajectories eligible for visual preference, consistent with the rise from 8.6% to 22.5% at step 124. The later decline below 1% is consistent with successful chains increasingly sharing the observer's preferred evidence style. Unlike caption-first warm-up in PRCO (Miao et al., 2026), this changing reward activity emerges without a scheduled change in image access.

The reward rule explains how such a pattern can arise: only correct chains scoring above the upper median receive a bonus, and four maximum-score ties among eight rollouts make $m = M$ and switch it off entirely. As successful chains converge in rating, the observer has fewer within-group distinctions to reward. Correctness gating and score ties both contribute to the logged activity, so the curve offers a plausible account of this shift rather than a direct measure of absolute fidelity. CoT length stays bounded through the change.

C Data and Evaluation Details

C.1 Bifocal-24K Construction

Bifocal-24K is designed as a common training substrate for studying the reward rather than a mixture specialized to one benchmark family. Each example is represented by an image, question, option set, and ground-truth option. We retain instances with valid images, unambiguous option mappings, and questions that require visual input. Source-specific schemas are converted to one response interface, while the original image and semantic content are preserved. We remove incompatible non-multiple-choice records and apply no synthetic augmentation. The resulting coverage ranges from scientific diagrams and textbook questions to scene understanding, charts, tables, and geometry.

Training examples and the evaluation splits used in this paper are disjoint, including where a training source and a benchmark share a dataset name.

This curation is a supporting contribution of the work: it makes the same RL recipe applicable across heterogeneous visual reasoning domains and reduces dependence on any single benchmark's answer format. The 19 sources span science (PMC-VQA, ScienceQA, AI2D, TQA, VQA-RAD, GEOS, EXAMS-V), general visual understanding (OpenSpaces, SpaceLLaVA, A-OKVQA, FigureQA, IconQA, MapQA, Visual7W), and multimodal mathematics (MultiMath-300K, TabMWP, UniGeo, Geometry3K, GeoQA+). Figure 3 visualizes their composition.

Source publications include PMC-VQA (Zhang et al., 2024b), ScienceQA (Lu et al., 2022), AI2D (Kembhavi et al., 2016), TQA (Kembhavi et al., 2017), VQA-RAD (Lau et al., 2018), GEOS (Seo et al., 2015), EXAMS-V (Das et al., 2024), A-OKVQA (Schwenk et al., 2022), FigureQA (Kahou et al., 2018), IconQA (Lu et al., 2021b), Visual7W (Zhu et al., 2016), MultiMath-300K (Peng et al., 2024), TabMWP (Lu et al., 2023), UniGeo (Chen et al., 2022), Geometry3K (Lu et al., 2021a), MapQA (Chang et al., 2022), and GeoQA+ (Cao & Xiao, 2022).

OpenSpaces and SpaceLLaVA (the source label `Spacellava`) are RemyxAI's VQASynth spatial-VQA releases (remyxai, 2024).[1]

C.2 Benchmarks and Baseline Groups

The evaluation suite includes: ChartQA (Masry et al., 2022), InfoVQA (Mathew et al., 2022), OCRBench (Liu et al., 2024b), MMMU (Yue et al., 2024), MMMU-Pro (Yue et al., 2025), MathVerse (Zhang et al., 2024a), MathVision (Wang et al., 2024), ScienceQA (Lu et al., 2022), AI2D (Kembhavi et al., 2016), MMStar (Chen et al., 2024), MMBench (Liu et al., 2024a), Real-

[1] `https://huggingface.co/datasets/remyxai/OpenSpaces`; `https://huggingface.co/datasets/remyxai/vqasynth_spacellava`.

WorldQA[2], HallusionBench (Guan et al., 2024), POPE (Li et al., 2023), and BLINK (Fu et al., 2024). Comparable-scale baselines cover outcome-driven RL and self-reflection (Vision-R1, VL-Rethinker, and MM-Eureka (Huang et al., 2026b; Wang et al., 2025; Meng et al., 2026)), description-based verification and coevolution (Vision-SR1, Visionary-R1, and PRCO (Li et al., 2026; Xia et al., 2026; Miao et al., 2026)), and perception-enhanced RL (PeBR-R1, Perceval, Perception-R1, VPPO, and PEARL (Chen et al., 2025b; Min et al., 2026; Xiao et al., 2026; Huang et al., 2026a; Zhang et al., 2026)).

C.3 Evaluation

Evaluation uses VLMEvalKit (Duan et al., 2024) across the 15 benchmarks. Base checkpoints use its default benchmark prompts. Reasoning baselines receive a concise, broadly shared instruction to reason before giving the final answer; Bifocal uses its trained reasoning format. Multiple-choice answers are first extracted by strict final-answer formatting rules. An auxiliary LLM checks item-level potential misses and false matches, while final scores follow each benchmark's native metric. Extraction and open-ended tasks retain the answer text required by that metric. OCRBench matches references against the complete saved prediction; VQA retains its task-specific soft accuracy or ANLS after answer extraction. No training-time verifier or visual scorer is used at test time.

Scores in Table 1 follow the metric native to each benchmark. HallusionBench reports aAcc. The Avg column is the arithmetic mean of the 15 benchmark scores displayed in Table 1, so every benchmark contributes once regardless of the task-family heading.

Prompt forms and answer fields. The evaluation input retains the benchmark image or images, question, and any supplied options or hints. Base checkpoints follow the benchmark's default answer instruction. For reasoning checkpoints, the common task requirement is to reason before returning a separately identifiable final answer. Bifocal uses the same `<think>` and `<answer>` fields as in training. Model-specific response interfaces, including a preceding description field for description-based methods, are retained where required by the corresponding wrapper; the common reasoning requirement does not imply an identical character-for-character prompt for every model.

For multiple-choice items, the final field contains the selected option label rather than the option's explanation. For binary questions, it contains Yes or No. For open-ended and extraction tasks, it contains the requested number, word, phrase, or mathematical answer in the benchmark's required form. These task-specific answer requirements do not supply intermediate observations or a reference solution. The evaluator reads the final answer for task scoring; the reasoning text is retained separately for chain analyses. The external output audit in Section C.4 evaluates this saved reasoning and does not alter the generated answer or provide a training reward.

C.4 VMCBench Chain Audit

For the four tasks in Table 4, we align the two 3B models' saved outputs with the same VMCBench-DEV items by question and index. Each task contributes all 50 DEV items. The four tasks were selected after examining answer accuracy; the 200-item result therefore characterizes these tasks, rather than estimating validity over all 1,000 DEV items.

Judge and information access. We call `qwen3.8-max-0902` through the DashScope-compatible multimodal API with thinking enabled and temperature zero. For each item, one image-conditioned request rates the two anonymized traces independently, with their display order randomized. Two separate text-only requests each assess one trace, preventing the other response from supplying missing premises. The requests contain the question, options, trace, and its stated answer, but not the gold label. Images are resized to a maximum edge of 1,280 pixels. Each judgment returns a score and a short supporting excerpt. Separate descriptions are included when present, but their facts count only if used or restated in the answer-producing reasoning.

Visual-support rating F. A score of 0 indicates a decisive visual claim contradicted by the image, including an incorrect number, text, object, attribute, or relation. A score of 1 indicates no concrete

[2] `https://x.ai/news/grok-1.5v`

answer-relevant visual claim, or insufficient image evidence to verify it. A score of 2 requires at least one decisive claim supported by the image and no decisive contradictory claim. Unrelated detail and greater length do not increase the rating.

Image-free sufficiency rating S. A score of 0 indicates no useful answer-relevant premise, a bare answer without an observation, or a contradictory trace. A score of 1 indicates useful content with a missing decisive fact or inference, or an unresolved material error. A score of 2 requires observations and, where necessary, reasoning sufficient to select the stated option without the image. A direct identification task can pass with one explicit observation; a multistep task must also supply the needed relation or calculation. The text-only judge treats stated visual observations as premises, while the separate F assessment checks their visual support. This is a rating of textual sufficiency, not a second rollout by the trained policy.

Endpoint and denominators. Let $c^{\text{exp}}_{m,i}$ indicate that model m gives an explicit final answer accepted by the benchmark, and let $F_{m,i}, S_{m,i} \in \{0, 1, 2\}$ be these audit ratings. Joint validity is

$$J_m = \frac{1}{200} \sum_{i=1}^{200} c^{\text{exp}}_{m,i} \mathbf{1}[F_{m,i} = 2] \mathbf{1}[S_{m,i} = 2]. \tag{23}$$

The score-bin distributions in Table 4(b) are marginals among explicitly correct chains; their intersection need not equal either marginal. The joint count is 115 of 140 explicitly correct Vision-SR1 chains and 126 of 158 Bifocal chains. Dividing by all 200 attempted items gives 57.5% and 63.0%. Dividing by explicitly correct answers instead gives 82.1% and 79.7%, a different question. The reported gain is therefore in valid solutions per attempted item, not the conditional validity of a correct answer. Task rows contribute to the same 200-item denominator; each task's answer accuracy uses its own 50 items.

Borderline judgments underwent item-level review under these criteria, and the table reports the final adjudicated counts. One benchmark-correct Vision-SR1 output lacks an unambiguous final label, explaining 141 benchmark-correct answers but 140 chains in the score distribution. All selected items remain in the denominator, including tasks solvable from textual context alone; such items can pass S without passing F if no decisive visual evidence is stated. This audit characterizes output quality on the selected tasks. Its external judge and rubric are distinct from the frozen training observer.

D Paired Diagnostic Protocols

These post-training diagnostics examine the information used by each training signal on Qwen2.5-VL-3B. We use the frozen pre-RL checkpoint for visual ratings and the trained Bifocal checkpoint for image-stripped recovery. Both use greedy decoding; the first-pass chains come from saved benchmark outputs. Visual ratings use the five-level evidence rubric in Appendix B.

Visual fact and image perturbations. Thirty synthetic OCR cards and thirty table-item images have known target values. For each image, we form two otherwise identical short chains that state the correct or an incorrect value drawn from another item of the same type. The question and image remain fixed. We also score the correct chain against a different image of the same type with a different target value, leaving its text unchanged. A separate exploratory set contains 18 AI2D and 20 MMStar first-pass-correct chains. Each has one occurrence of a numeric gold-option value that can be replaced by a different numeric option of the same character length; all other chain text is unchanged. A correct final answer does not certify every statement in the original chain, so this set tests score sensitivity rather than calibrated claim accuracy.

Table 3(a) reports the paired counts, with the difference defined as the original rating minus the perturbed rating. The controlled value-substitution difference averages 3.53 rating points (paired bootstrap 95% interval 3.30–3.75); the real-CoT difference averages -0.03 (-0.24–0.13). Most real chains retain the top rating after one value changes. Together with the image-swap results, this identifies the score's operating granularity: it distinguishes image-conditioned evidence at the whole-chain level, with limited sensitivity to isolated edits inside an otherwise detailed trace.

Chain removal and mismatch. We select 40 saved responses from each of AI2D_TEST, MM-Star, and MMMU_DEV_VAL, balancing 20 initially correct and 20 initially incorrect answers per benchmark. For each item, a fixed cyclic option permutation and its remapped gold label are shared across the three conditions. The trained policy receives no image and sees either its original answer-producing chain, no chain, or the chain of another selected item from the same benchmark. The question and permuted options are identical across conditions. All 360 second-pass outputs are generated greedily; malformed labels count as incorrect.

Table 3(b) gives the recovery counts. The original chain exceeds the empty and mismatched conditions by 21.7 and 22.5 percentage points over all items (paired bootstrap 95% intervals 11.7–31.7 and 11.7–33.3 points). The gain is clearest when the first answer was correct. The comparison establishes that the chain contributes answer-relevant information beyond the shared question and options. The image-conditioned test separately examines visual support for its premises, completing the two-view diagnostic.

For a comparison between conditions A and B, let $r_i^A, r_i^B \in \{0, 1\}$ indicate recovery success on item i. The reported recovery difference is $100n^{-1}\sum_i(r_i^A - r_i^B)$ percentage points. Paired bootstrap resampling keeps the two outcomes for an item together; visual-score differences are treated in the same paired manner. This preserves the matched-item structure of the interventions.

D.1 Reading the Component Comparisons

Table 2 separates three choices. Full Bifocal versus w/o Visual tests the added visual branch with a co-evolving sufficiency verifier. Full Bifocal versus Co-ev. Fidelity compares frozen and co-evolving visual observers under that verifier. Finally, w/o Visual versus Frozen Suf. compares the sufficiency checkpoints without the visual branch. Together, these comparisons examine visual anchoring and the asymmetric checkpoint choice in the evaluated settings.

R-Avg and E-Avg summarize the reasoning and extraction tasks defined in Table 2. The larger extraction gain after adding the visual branch is consistent with preserving answer-relevant visual details. These component comparisons examine the added branch and checkpoint choices, while the paired diagnostics examine the signals and the VMCBench audit examines the resulting output chains. They evaluate complementary aspects of the design rather than a full factorial interaction between the rewards.

E Visual-Evidence Diagnostic Details

E.1 Local Occlusion Sensitivity

For an image V and gold answer a^*, let $\ell_m(V; a^*)$ be model m's mean teacher-forced log-probability over the non-special answer tokens. We divide the image into a 6×8 grid, blur one cell c at a time to obtain $\widetilde{V}_c$, and define

$$\begin{aligned} \ell_m(V; a^*) &= \frac{1}{|\mathcal{T}(a^*)|} \sum_{j \in \mathcal{T}(a^*)} \log p_m\big(a_j^* \mid V, Q, a_{<j}^*\big)\,, \\ s_{m,c} &= \ell_m(V; a^*) - \ell_m(\widetilde{V}_c; a^*), \\ S_m(R) &= \sum_{c \in R} [s_{m,c}]_+, \qquad C_m(R) = \frac{S_m(R)}{\sum_c [s_{m,c}]_+}. \end{aligned} \tag{24}$$

Here $[x]_+ = \max(x, 0)$; when the total positive sensitivity is zero, we set $C_m(R) = 0$. $S_m(R)$ measures the absolute positive sensitivity inside a manually marked answer-relevant region R, while $C_m(R)$ measures how much of the image-wide positive sensitivity is concentrated there. All four checkpoints use the same processor, image resolution, Gaussian blur, grid, gold target, and region; the heatmaps share a color scale within each case. The regions are chosen from the question and visible evidence and are not optimized separately for a model. These are case-level interventions, not estimates of dataset-wide causal effects.

Figures 7–11 add five cases selected from a broader output screen. The first three show distinct Bifocal corrections. In AI2D-3014 and ChartQA-1211, Bifocal is the only model to answer correctly

and has the largest $S_m(R)$. In OCRBench-0500, it is the only model to retrieve the correct value and has the highest $C_m(R)$, although two baselines have larger absolute sensitivity. The remaining AI2D examples show a different regime: Bifocal's total relevant-region sensitivity is modest, but its concentration is highest and its decision is correct. Thus neither larger $S_m(R)$ nor larger $C_m(R)$ alone is a correctness certificate. The consistent observation is that evidence must be both selected and used by the subsequent reasoning.

E.2 Depth-Wise Visual-Attention Mass

Let $A^{m,l}_{h,q,k}$ be attention from query token q to key token k at head h and layer l. For query set $\mathcal{Q}_r$, image-token set $\mathcal{I}$, and H heads, we compute

$$
\begin{aligned}
v^{(r)}_{m,l} &= \frac{1}{H|\mathcal{Q}_r|} \sum_{h=1}^{H} \sum_{q\in\mathcal{Q}_r} \sum_{k\in\mathcal{I}} A^{m,l}_{h,q,k}, \\
\Delta^{(r)}_{m,l} &= \operatorname{logit}\Big(v^{(r)}_{m,l}\Big) - \operatorname{logit}\Big(v^{(r)}_{\pi_0,l}\Big),
\end{aligned}
\tag{25}
$$

where $\operatorname{logit}(x) = \log \frac{x}{1-x}$; values are clipped to $[10^{-8}, 1-10^{-8}]$ before the transform. Prompt-side queries are non-image tokens in the user turn; answer-side queries are teacher-forced gold-answer tokens. The probe contains 992 fixed inputs from AI2D, BLINK, ChartQA, HallusionBench, MathVision, MMStar, OCRBench, POPE, and RealWorldQA. Every checkpoint receives identical image, prompt, and answer-token sequences. The main-paper curves use a three-layer moving average and 95% intervals from 2,000 paired bootstrap resamples over examples.

Figure 12 averages the same change within early, middle, and late layer blocks for each benchmark. The aggregate pattern is not attributable to one task: Vision-SR1 usually reduces early and middle mass, PRCO generally increases it, and Bifocal remains close to the initialization on prompt-side queries before reducing late answer-side mass. Task-specific late-layer deviations remain, so we interpret the result as a descriptive redistribution rather than a monotone measure of grounding quality or a causal explanation of the reward.

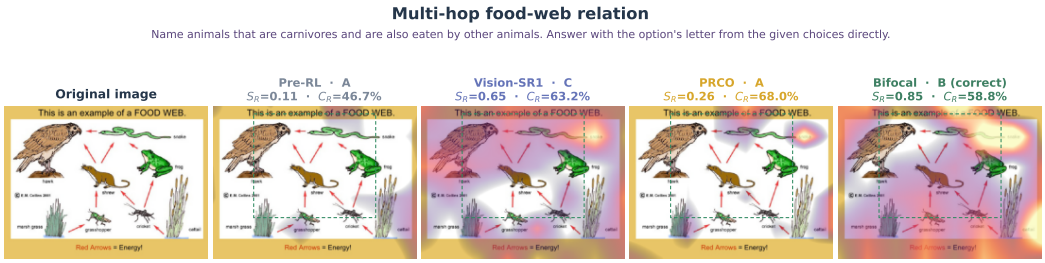


Figure 7: Multi-hop food-web reasoning. Bifocal is the only model to select the shrew and has the largest answer-relevant absolute sensitivity ($S_R = 0.85$).



Figure 8: Chart extraction followed by arithmetic. Bifocal alone extracts the two relevant values and returns their correct sum, 399,075; its $S_R = 6.94$ is the largest among the four models.

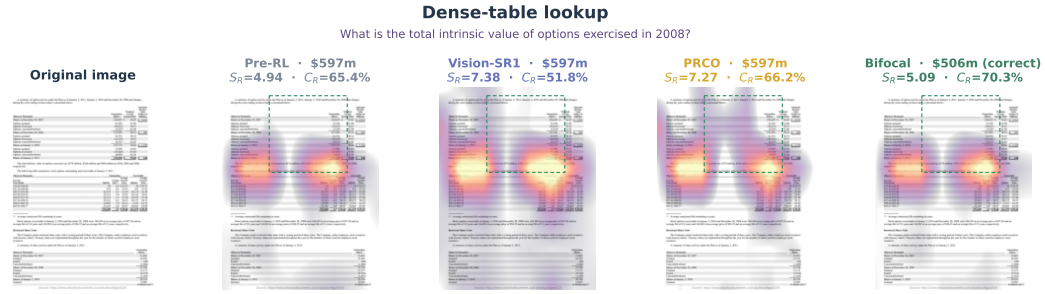


Figure 9: Dense-table lookup. The three comparison models return $597 million, whereas Bifocal retrieves $506 million and has the highest sensitivity concentration in the marked table region ($C_R = 70.3\%$).

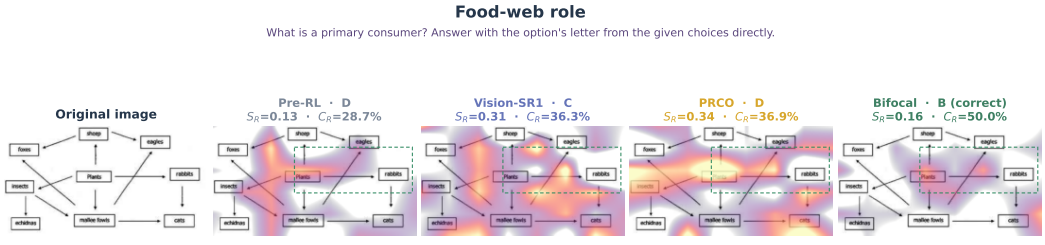


Figure 10: Food-web role recognition. Bifocal is the only model to identify the rabbit as the primary consumer; its positive sensitivity is most concentrated in the marked relation ($C_R = 50.0\%$).

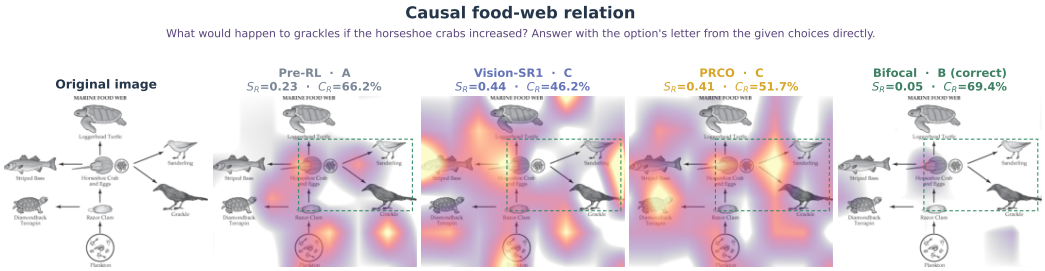


Figure 11: Causal food-web reasoning. Bifocal is the only model to predict that grackles increase. Its absolute sensitivity is small, but the largest share falls on the horseshoe-crab–grackle relation ($C_R = 69.4\%$), illustrating why concentration and magnitude must be interpreted together.

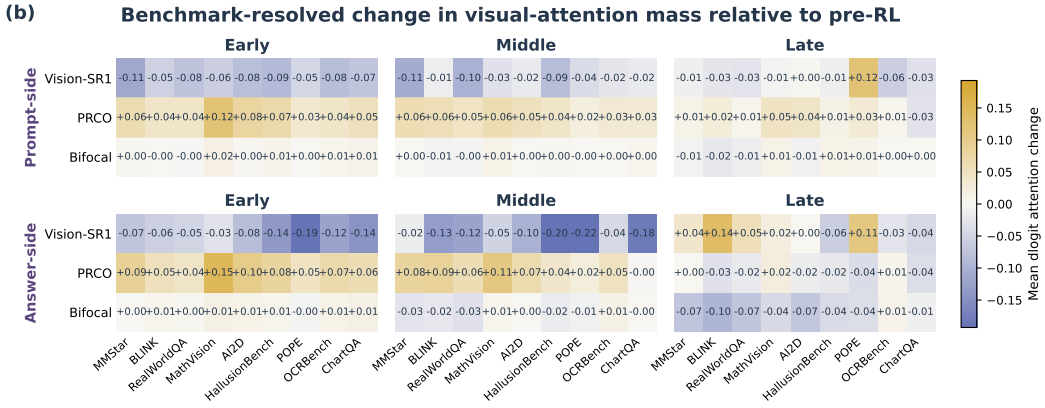


Figure 12: Benchmark-resolved change in visual-attention mass relative to the common pre-RL initialization. Each cell averages $\Delta^{(r)}_{m,l}$ within an early, middle, or late layer block.

F Extended Related Work

F.1 Reinforcement Learning for VLM Reasoning

R1-VL (Zhang et al., 2025a) develops step-wise GRPO, Visual-RFT (Liu et al., 2025) applies reinforcement fine-tuning to visual tasks, and Vision-R1 (Huang et al., 2026b) combines a cold start with RL. VLM-R1 (Shen et al., 2025) studies stable, generalizable multimodal RL, while VLAA-Thinker (Chen et al., 2025a) compares SFT and RL for visual reasoning. MM-Eureka (Meng et al., 2026) scales rule-based outcome RL for multimodal math. Most retain answer correctness as their primary signal and leave the visual basis of a successful trajectory implicit.

F.2 Self-Verification and Self-Rewarding

Vision-SR1 (Li et al., 2026) and Visionary-R1 (Xia et al., 2026) verify a separately generated description layer rather than the think chain that drives the answer, leaving a gap between what is verified and what is used. PRCO (Miao et al., 2026) alternates a shared trainable policy between an Observer that writes question-conditioned evidence captions and a Solver whose downstream success supplies the caption utility reward, with an auxiliary LLM suppressing answer leakage. This coevolution makes the intermediate description more useful, but the rewarded object remains a caption proxy and utility does not directly test whether its visual claims are supported by the image. ACRE (Huang et al., 2025) verifies the think chain directly but retains the image, so its consistency test checks option-order invariance rather than dependence on visual evidence. These methods re-evaluate intermediate traces in different ways, but none jointly applies a text-only sufficiency test and an image-conditioned visual-evidence test to the same answer-producing chain.

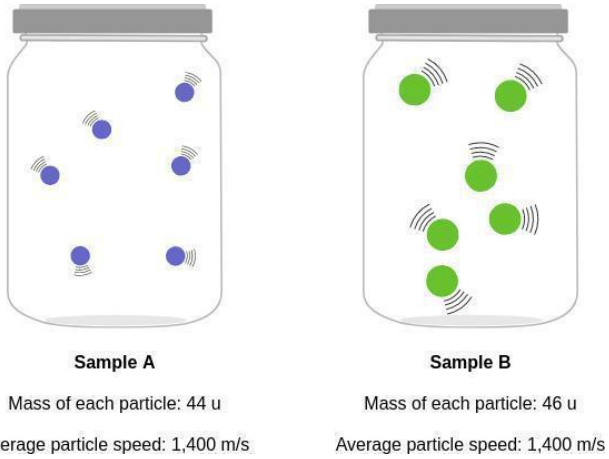
F.3 Perception Enhancement and Visual Evidence

Perception-enhanced RL methods introduce visual supervision through several routes. Perception-R1 (Xiao et al., 2026) uses a text-only LLM judge to check whether responses reflect visual key information extracted from teacher-generated trajectories; PeBR-R1 (Chen et al., 2025b) rewards image descriptions through FGCLIP alignment and teacher-derived keywords before reasoning RL; Perceval (Min et al., 2026) uses a frozen process reward model for token-level hallucination detection; PEARL (Zhang et al., 2026) constructs perception checklists whose auxiliary rewards also gate reasoning updates. VPPO (Huang et al., 2026a) computes a visual-dependency signal from the policy's own distribution change under image masking and uses it to shape the advantage and filter gradients. These approaches rely on annotations, external representation spaces, auxiliary evaluators, or counterfactual masking. Bifocal instead turns the policy's pre-RL visual competence into a trajectory-level anchor and couples it to sufficiency verification on the same chain (Section 3).

G Complete Bifocal Reasoning Examples

Figures 13–15 present complete saved Qwen2.5-VL-3B Bifocal responses from three tasks. Each pairs the original benchmark image and question with the full think-field text and final answer, preserving the generated wording and changing only typesetting. The examples were selected by checking the stated observations against the image and following the inference to the benchmark answer. They illustrate scientific inference, chart arithmetic, and diagram-based identification. In each, the visual premise is expressed in the same chain that uses it to answer.

MMStar, item 619. Compare the average kinetic energies of the particles in each sample. Which sample has the higher temperature?
A. Neither; the samples have the same temperature. B. Sample B. C. Sample A.
Ground truth: B (sample B).



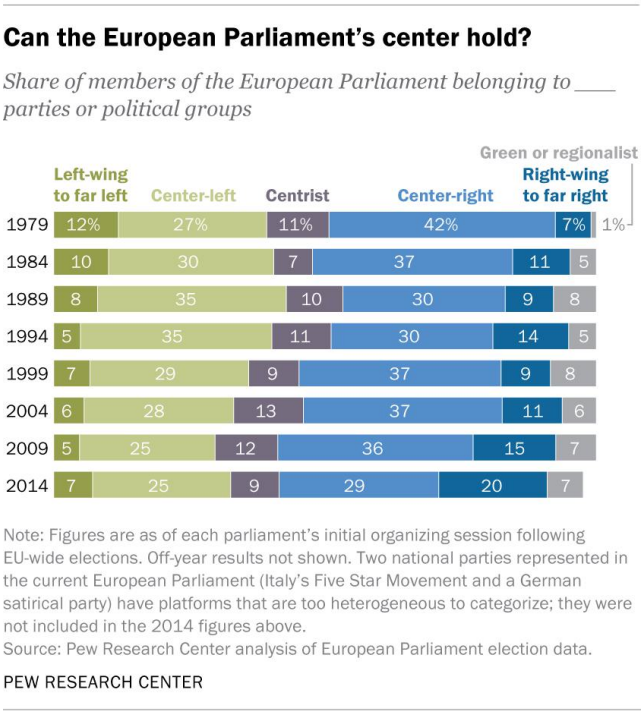
Complete CoT.

> Let's analyze the image and question carefully. The image shows two samples, A and B, with particles of different masses but the same average speed. The key information is that the particles in sample B have a higher mass than those in sample A, but they move at the same speed. In physics, temperature is directly related to the average kinetic energy of particles, which is proportional to the square of their speed. Since the particles in sample B have a higher mass, they have a higher average kinetic energy and thus a higher temperature. The image explicitly shows this difference in mass and speed, making it clear that sample B has the higher temperature. The reasoning is based on the direct relationship between particle mass and temperature in the context of the image.

Generated answer: B.
Case analysis. The image gives masses of 44 and 46 u at the same average speed, 1,400 m/s. Under the particle comparison specified by the question, the larger mass implies larger kinetic energy at that speed. The chain preserves this visual comparison and uses it to select sample B.

Figure 13: Complete scientific-reasoning response: a visual comparison supplies the premise for the physical inference.

ChartQA, item 74. What is the average of the smallest gray bar and largest light blue bar?
Ground truth: 21.5.



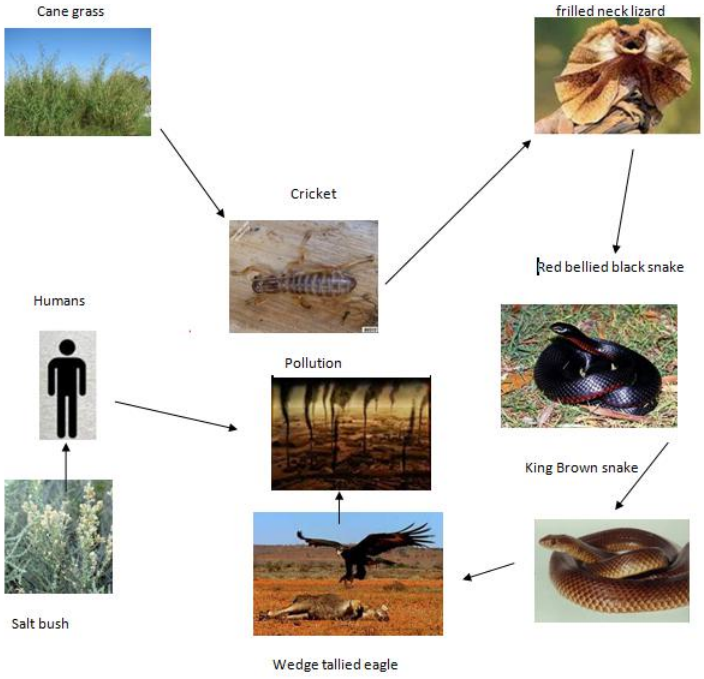
Complete CoT.

> Let's carefully analyze the image to find the smallest gray bar and the largest light blue bar. The smallest gray bar is 1%, and the largest light blue bar is 42%. The image clearly shows these values. To find the average, we add these two numbers (1% + 42%) and divide by 2. The calculation is straightforward: (1 + 42) / 2 = 43 / 2 = 21.5. The image directly provides the answer, making the calculation simple and direct. The reasoning is based on the visual data clearly presented in the image. No additional calculations are needed beyond this straightforward addition and division.

Generated answer: 21.5.
Case analysis. The gray minimum is 1 and the light-blue maximum is 42. Both extracted values are carried into the stated calculation, $(1 + 42)/2 = 21.5$, without changing their numerical values or roles.

Figure 14: Complete chart-reasoning response: the selected visual extrema are preserved in the arithmetic that produces the answer.

AI2D, item 188003. What feeds on crickets?
A. Snake. B. Lizards. C. Eagle. D. Salt bush.
Ground truth: B (lizards).



Complete CoT.

> Let's analyze the image step-by-step to determine what feeds on crickets. The image clearly shows a direct arrow from crickets to a frilled neck lizard, indicating that the frilled neck lizard feeds on crickets. This is explicitly shown in the diagram, making it a straightforward and accurate representation of the food chain. The other options like snakes, eagles, or salt bush are not shown as feeding on crickets in the image. The image provides a clear, visual representation of the food chain, making the answer self-evident. The reasoning is based on the direct visual information provided in the image.

Generated answer: B.
Case analysis. The arrow from cricket to frilled neck lizard establishes the queried feeding relation. The snake appears downstream of the lizard, so the chain uses the direct relation relevant to the question to select the lizard option.

Figure 15: Complete diagram-reasoning response: the chain states the relevant directed relation and maps it to the answer option.