REINFORCEMENT LEARNING WITH VISUALLY ATTRIBUTABLE ON-POLICY SELF-DISTILLATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Multimodal reinforcement learning with verifiable rewards (RLVR) provides reliable task-level supervision, yet offers limited guidance for localizing errors in visual reasoning. On-policy self-distillation (OPSD) complements this supervision with dense token-level corrections from a privileged teacher equipped with an evidence-centered visual view. However, the resulting corrections may conflate visual evidence with non-visual shifts induced by linguistic priors, response style, or the privileged view itself. We formalize this phenomenon as *privileged-correction contamination*: a token-level correction is contaminated when its update direction cannot be consistently attributed to the privileged visual evidence. Directly distilling contaminated corrections may therefore reinforce erroneous token-level updates. To address this issue, we propose reinforcement learning with visually attributable on-policy self-distillation, an evidence-consistent paradigm that augments task-level RLVR supervision with privileged token-level corrections grounded in visual evidence. We instantiate this paradigm with **Visual-Attribution Gating (VAG)**, which uses the verifier-derived advantage as the optimization anchor and validates each privileged correction against the contrast between evidence-present and evidence-degraded teacher views. By admitting only corrections directionally supported by this visual contrast, VAG preserves visually grounded guidance while suppressing inconsistent residuals. Experiments across six visual reasoning benchmarks show that VAG consistently outperforms GRPO and OPSD methods, establishing new state-of-the-art average performance.

1 INTRODUCTION

Reinforcement learning with verifiable rewards (RLVR) has become an effective paradigm for improving multimodal reasoning by optimizing policies against automatically checkable outcomes (Guo et al., 2025; Yang et al., 2025). In GRPO (Shao et al., 2024), a verifier scores each rollout and group-relative normalization turns these rewards into an outcome advantage A_{ORM} shared by every token. Such supervision is reliable at the trajectory level, but coarse at the token level: the same advantage is assigned to every token, regardless of which visual decision determines the answer. Consequently, RLVR can distinguish successful trajectories from unsuccessful ones without localizing the visual reasoning errors within them.

On-policy distillation (OPD) (Agarwal et al., 2024; Gu et al., 2023; Song & Zheng, 2026) addresses this limitation by evaluating student-sampled trajectories with a teacher and supplying dense token-level guidance. On-policy self-distillation (OPSD) (Zhao et al., 2026; Hübotter et al., 2026; Yang et al., 2026) further avoids dependence on an external teacher by using the policy itself under privileged information as the teacher, which for multimodal reasoning can be a zoomed-in view that exposes task-relevant evidence (Yuan et al., 2026). The resulting teacher–student log-probability gap δ_t provides a dense, token-level correction, yet it is not guaranteed to be attributable to the privileged visual evidence alone: it may also be contaminated by non-visual shifts induced by linguistic priors, response style, or the privileged view itself.

We formalize this contamination as *privileged-correction contamination*: the teacher–student correction mixes evidence-grounded guidance with residual effects that are not attributable to the visual evidence. Applying such a correction directly may reinforce a token even when the visual evidence argues against it, or suppress a token that the evidence supports.

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

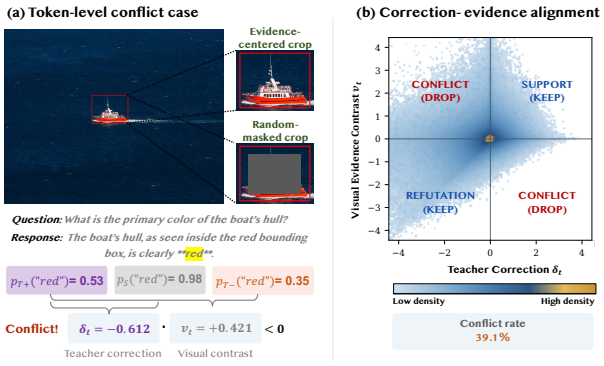


Figure 1: Illustrative contamination diagnostic. (a) A token-level conflict case and (b) the correction-evidence alignment used to motivate attribution.

To make this contamination concrete, Fig. 1 illustrates a single token-level diagnostic. For each generated token, we compare the teacher correction δ_t with the visual contrast v_t between evidence-centered and random-masked teacher views, a probe of visual dependence introduced by OPD-V (Aniri et al., 2026). Fig. 1 (a) shows a conflicting case, in which the teacher suppresses a token ($\delta_t = -0.612$) that the visual contrast supports ($v_t = +0.421$). Such a correction should be discarded rather than applied, since the student already assigns the token a higher probability than the teacher does (0.98 vs. 0.53): the teacher is lower here because the privileged view shifts its probabilities downward, not because the visual evidence refutes the token. Fig. 1 (b) summarizes the comparison over tokens by the signs of the two signals: agreeing signs indicate a correction consistent with the evidence, whereas opposite signs expose the two conflict cases, in which the teacher suppresses a visually supported token or promotes a visually refuted one. Measured with Qwen3.5-4B on our training data (§4.1), these conflicts affect 39.1% of the generated tokens. Exposing these conflicts, however, does not remove them. What the diagnostic calls for is visual attribution in a strict sense: an explicit mechanism that checks whether the *direction* of each correction agrees with the visual evidence.

Motivated by this diagnosis, we design **Visual-Attribution Gating (VAG)**. Rather than distilling every privileged correction, VAG validates the direction of each one against the visual contrast: the correction δ_t is admitted only when it agrees in sign with v_t, that is $v_t\delta_t > 0$, which keeps both supporting and refuting corrections and removes only the conflicting residuals exposed by the diagnostic. Furthermore, VAG operates on top of RLVR rather than in place of it: the verifier-derived advantage remains the optimization anchor, and self-distillation contributes only the corrections that pass the gate. Taken together, we propose reinforcement learning with visually attributable on-policy self-distillation, a new evidence-consistent learning paradigm.

Fig. 2 compares our paradigm with representative on-policy learning paradigms for multimodal credit assignment along three properties: *trajectory correctness*, *dense token correction*, and *visual attribution*. RLVR (Fig. 2 (a)) supplies only the outcome advantage A_{ORM}, which anchors optimization to trajectory correctness but assigns the same scalar to every token. RLVR with distillation (Fig. 2 (b)) keeps this anchor and modulates each token by δ_t without verifying where that correction comes from, so the privileged signal it distills may be contaminated (Lin et al., 2026; Yang et al., 2026). Visual contrastive distillation (Fig. 2 (c)) instead uses visual contrast v_t to select only visually sensitive tokens (Aniri et al., 2026). Yet this selection does not verify whether an update follows the visual evidence, so the corrections applied to the selected tokens remain contaminated, while the verifier-grounded anchor is lost as well.

VAG (Fig. 2 (d)) is the only paradigm that attains all three properties at once: it keeps A_{ORM} as the anchor, preserves the dense correction δ_t, and admits that correction only when v_t supports its direction. This coverage comes at almost no cost. VAG needs no learned contamination estimator and no auxiliary distillation objective, and relative to a dual-view OPSD baseline that already evaluates both teacher views, it adds only an elementwise sign test and no additional teacher forward pass. Across six visual reasoning benchmarks, VAG consistently improves over matched GRPO and OPSD baselines and surpasses much larger open-source models (e.g., Qwen3.5-397B), leading closed-source models (e.g., GPT-5.4), and agentic "Thinking-with-Images" methods on tasks that require dense local evidence. On holdout tasks beyond the training distribution, it retains general visual understanding and reasoning ability, indicating that the gains do not come at the cost of forgetting, while controlled ablations isolate the contribution of the sign-agreement gate.

Our contributions are summarized as follows:

- We identify *privileged-correction contamination*, showing that dense teacher–student corrections are not necessarily aligned in direction with the visual evidence.

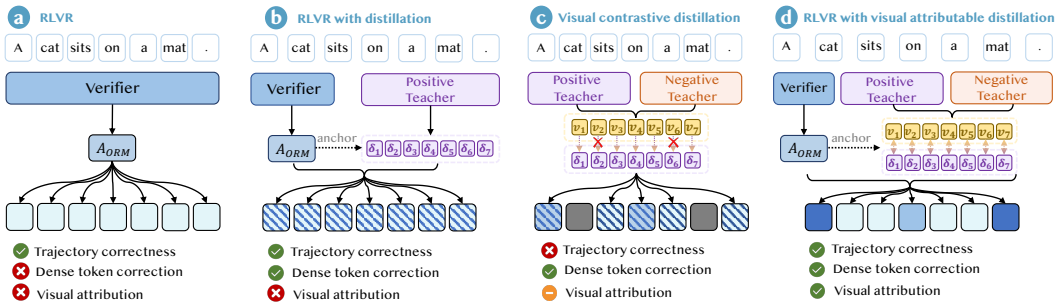


Figure 2: **Comparison of on-policy learning paradigms.** (a) RLVR, (b) RLVR with distillation, (c) Visual contrastive distillation, and (d) RLVR with visual attributable distillation.

- We propose an evidence-consistent paradigm unifying RLVR and OPSD, in which the verifier-derived advantage anchors optimization, and design VAG to mitigate privileged-correction contamination by admitting a teacher correction only under sign agreement with a controlled visual contrast.
- Across six visual reasoning benchmarks, VAG consistently improves over matched GRPO and OPSD baselines and surpasses much larger open-source and leading closed-source models, while holdout evaluation shows no loss of general capability. Ablations further validate the proposed attribution mechanism.

2 Related Work

2.1 On-Policy and Privileged Distillation

OPD supervises student-visited prefixes with a teacher, thereby reducing exposure bias while supplying dense token-level guidance (Gu et al., 2023; Agarwal et al., 2024; Song & Zheng, 2026). OPSD replaces the external teacher with the model itself conditioned on privileged context, such as verified solutions, reasoning traces, or environmental feedback (Zhao et al., 2026; Hübotter et al., 2026; He et al., 2026; Jin et al., 2026). Because the student cannot access this context at inference time, the resulting teacher–student gap may conflate task-relevant knowledge with leakage or stylistic artifacts (Yang et al., 2026; Pan et al., 2026; Shen et al., 2026). In multimodal reasoning, privileged views instead expose localized evidence that global image processing tends to lose, and methods from region-to-image distillation to Vision-OPD transfer such evidence to a full-image model (Wei et al., 2026; Yuan et al., 2026). Subsequent work refines the selectivity of this supervision through view-based reweighting (Liu et al., 2026), trajectory gating (Sun et al., 2026), gradient decomposition (Yoon et al., 2026), modality balance (Aniri et al., 2026), and target reconstruction (Zhang et al., 2026), whereas VAG uses a signed visual contrast to determine whether a candidate teacher correction is admitted into policy optimization at all.

2.2 Integrating RLVR with On-Policy Distillation

RLVR anchors optimization to checkable outcomes, and GRPO realizes this without a learned critic by estimating group-relative advantages that are nevertheless shared by all tokens in a response (Shao et al., 2024). OPD supplies the missing token-level resolution, yet its preference need not agree with trajectory correctness, so the two signals are complementary but not automatically compatible. Existing integrations reconcile them by weighting the two objectives, restricting the teacher gap to modulating verifier-directed magnitudes, gating implicit OPD rewards by trajectory correctness, or contrasting privileged contexts (Hübotter et al., 2026; Yang et al., 2026; Lin et al., 2026; Pan et al., 2026), while variants route samples or tokens between the two objectives, account for student uncertainty, or reverse teacher signals to encourage exploration (Li et al., 2026; Wang et al., 2026; Ke et al., 2026; Kim et al., 2026). VAG resolves this compatibility at the token level instead: a signed correction enters the verifier-anchored advantage only when its direction agrees with the visual evidence contrast, which preserves both visually supported promotion and visually justified suppression while falling back to outcome supervision whenever attribution fails.

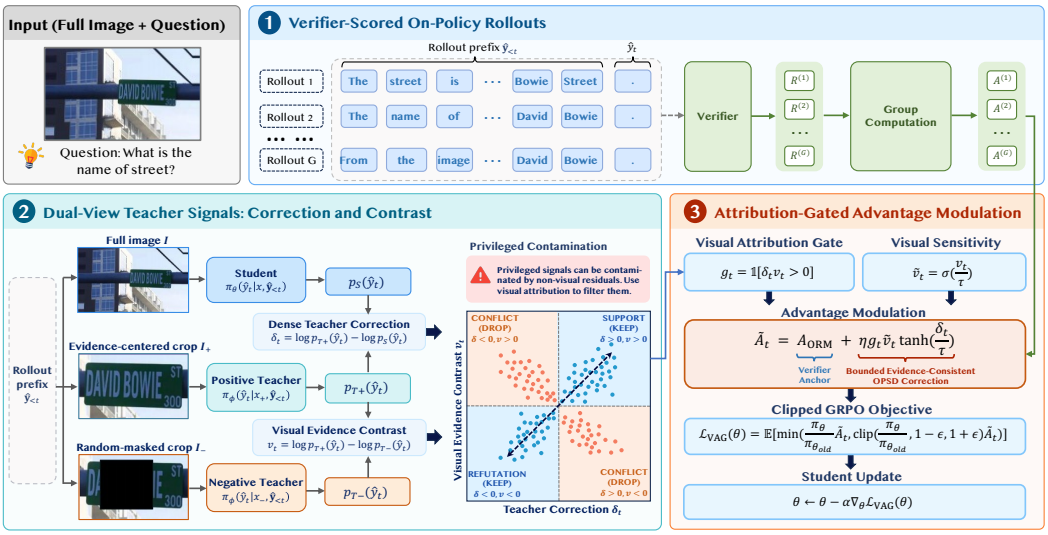


Figure 3: Overview of the VAG framework. **Stage 1:** A verifier derives A_{ORM} from on-policy rollouts. **Stage 2:** Dual-view teacher evaluation yields the distillation margin δ_t and visual contrast v_t; VAG retains same-sign corrections and discards conflicting ones. **Stage 3:** The retained correction modulates A_{ORM} for the clipped GRPO update.

3 Method

3.1 Overview

Given an image–question pair, VAG proceeds through the three-stage training loop illustrated in Fig. 3. In Stage 1 (Sec. 3.2), VAG derives a reliable trajectory-level advantage from on-policy rollouts sampled by the frozen student policy and scored by a task verifier. In Stage 2 (Sec. 3.3), the positive and negative teacher branches of the same Exponential Moving Average (EMA) model replay each student-generated prefix under an evidence-centered view and its random-masked counterpart, respectively. These evaluations yield a teacher–student correction δ_t and a visual contrast v_t, whose sign agreement retains evidence-consistent token-level supervision while filtering potentially contaminated corrections. In Stage 3 (Sec. 3.4), VAG weights each retained correction by its visual sensitivity and uses it to modulate, rather than replace, the verifier advantage. The resulting token-level advantage is then optimized using the clipped GRPO objective.

3.2 Verifier-Scored On-Policy Rollouts

Let q denote a question and I its corresponding full image, with the multimodal input denoted by $x = (q, I)$. At each training iteration, π_θ denotes the student policy used for rollout collection. As shown in Stage 1 of Fig. 3, the student samples a group of G response trajectories as:

$$\hat{\mathbf{y}}^{(i)} = \left(\hat{y}_1^{(i)}, \ldots, \hat{y}_{T_i}^{(i)}\right) \sim \pi_\theta(\cdot \mid x), \qquad i = 1, \ldots, G, \tag{1}$$

where $\hat{\mathbf{y}}^{(i)}$ is the i-th sampled response, $\hat{y}_t^{(i)}$ is its token at position t, and T_i is its length. A task verifier assigns each rollout an outcome reward $R^{(i)} \in \{0, 1\}$, from which we compute the group-normalized advantage as:

$$A_{\text{ORM}}^{(i)} = \frac{R^{(i)} - \mu_G}{\sigma_G}, \tag{2}$$

where μ_G and σ_G denote the mean and standard deviation of the rewards within the rollout group, respectively. The scalar $A_{\text{ORM}}^{(i)}$ is shared by all tokens in rollout i. Although it provides a reliable trajectory-level learning signal, it does not identify which tokens contribute to the outcome.

3.3 Dual-View Teacher Signals: Correction and Contrast

Stage 1 provides a reliable outcome-level anchor but leaves token-level credit undetermined. OPSD closes this gap by letting a privileged teacher re-score the student's own rollout token by token (Zhao

et al., 2026), and we adopt this mechanism as our dense signal, instantiating the teacher with an evidence-centered view. Unlike prior OPSD, which distills every correction directly, we first ask whether a correction is directionally supported by the visual evidence, following the diagnostic in Sec. 1. As shown in Stage 2 of Fig. 3, both teachers are view-conditioned branches of a single EMA teacher π_ϕ replaying the student's trajectory: conditioning π_ϕ on the evidence-preserving view yields the *positive teacher*, and conditioning it on the evidence-degraded view yields the *negative teacher*. Because the two branches share the student-generated prefix and differ only in the visual condition, all token-level quantities below are comparable. This yields two steps: form the privileged correction from the positive teacher, then probe its visual support with the negative teacher.

Dense Teacher Correction. The first step instantiates the positive branch. The evidence-preserving view I_+ is a crop centered on the target detail relevant to the question, which gives the positive teacher input $x_+ = (q, I_+)$. For each student-generated trajectory $\hat{\mathbf{y}}^{(i)}$, the positive teacher re-evaluates every generated token $\hat{y}_t^{(i)}$ conditioned on the same student-generated prefix $\hat{\mathbf{y}}_{<t}^{(i)} = (\hat{y}_1^{(i)}, \ldots, \hat{y}_{t-1}^{(i)})$. We write the corresponding sampled-token probabilities as:

$$p_S^{(i)}(\hat{y}_t) \triangleq \pi_\theta\left(\hat{y}_t^{(i)} \mid x, \hat{\mathbf{y}}_{<t}^{(i)}\right), \tag{3}$$

$$p_{T+}^{(i)}(\hat{y}_t) \triangleq \pi_\phi\left(\hat{y}_t^{(i)} \mid x_+, \hat{\mathbf{y}}_{<t}^{(i)}\right), \tag{4}$$

for the student and the positive teacher, respectively. The dense teacher correction is their log-ratio on the sampled token:

$$\delta_t^{(i)} = \mathrm{sg}\left[\log p_{T+}^{(i)}(\hat{y}_t) - \log p_S^{(i)}(\hat{y}_t)\right], \tag{5}$$

where $\mathrm{sg}[\cdot]$ denotes the stop-gradient operator. A positive $\delta_t^{(i)}$ indicates that the positive teacher assigns the sampled token a higher probability than the student, and a negative value the opposite, so $\delta_t^{(i)}$ is precisely the dense signal that on-policy self-distillation would follow.

Taken on its own, however, $\delta_t^{(i)}$ is exactly the quantity examined in Fig. 1: a non-zero teacher–student gap cannot be assumed to follow the direction of the privileged evidence. We therefore treat $\delta_t^{(i)}$ as a candidate correction and, instead of assuming its visual support, make that support testable.

Visual Evidence Contrast. Following OPD-V (Aniri et al., 2026), we pair I_+ with an evidence-degraded counterpart I_-, obtained by occluding a randomly placed rectangular region of the zoom-in image with black pixels, so that the visual evidence available to the teacher is weakened while the image resolution, the visual token grid, and the surrounding spatial context remain unchanged. Feeding $x_- = (q, I_-)$ to the same teacher on the same prefix gives:

$$p_{T-}^{(i)}(\hat{y}_t) \triangleq \pi_\phi\left(\hat{y}_t^{(i)} \mid x_-, \hat{\mathbf{y}}_{<t}^{(i)}\right), \tag{6}$$

and because the two teacher evaluations share parameters, prefix, and realized token, and differ only in the target evidence, their log-ratio isolates the teacher's response to that intervention:

$$v_t^{(i)} = \mathrm{sg}\left[\log p_{T+}^{(i)}(\hat{y}_t) - \log p_{T-}^{(i)}(\hat{y}_t)\right]. \tag{7}$$

A positive $v_t^{(i)}$ indicates that the target evidence supports the sampled token, whereas a negative value indicates refutation, and $|v_t^{(i)}|$ measures how strongly the token depends on that evidence.

Diagnosing Privileged Contamination. Pairing the two signals by sign partitions the generated tokens into the four quadrants shown in Stage 2 of Fig. 3. The two same-sign quadrants are directionally consistent: when $\delta_t > 0$ and $v_t > 0$, the teacher promotes a token whose probability the visual evidence also raises (*support*); when $\delta_t < 0$ and $v_t < 0$, the teacher suppresses a token whose probability the evidence lowers (*refutation*). In both cases the candidate correction moves the student in the same direction as the visual intervention, so it can be treated as evidence-grounded credit. Notably, keeping the refutation quadrant is what allows suppressive corrections to survive rather than be discarded together with the conflicting ones.

The two opposite-sign quadrants are directionally conflicting. A correction with $\delta_t > 0$ but $v_t < 0$ would reinforce a token whose probability the visual evidence lowers, whereas $\delta_t < 0$ with $v_t > 0$ would suppress a token whose probability it raises. Since neither direction can be explained by the visual intervention, we regard these quadrants as observable instances of privileged-correction contamination and exclude them from the dense supervision. The four quadrants thus define a binary admissibility rule on candidate corrections, and what remains is to realize this rule as a token-level learning signal without discarding the verifier anchor of Stage 1.

3.4 Attribution-Gated Advantage Modulation

Stage 3 realizes the admissibility rule as a token-level advantage: the gate decides *whether* a candidate correction enters the update, the contrast magnitude decides *how strongly* it enters, and the outcome advantage of Stage 1 remains the anchor that determines the direction of the update. The visual attribution gate shown in Fig. 3 is formulated as:

$$g_t^{(i)} = \mathbb{1}\Big[\delta_t^{(i)} v_t^{(i)} > 0\Big], \tag{8}$$

where $\mathbb{1}[\cdot]$ is the indicator function. This gate retains the support and refutation quadrants and rejects both conflict quadrants, together with the boundary cases that carry no directional information. We therefore convert the contrast into a bounded visual-sensitivity weight:

$$\widetilde{v}_t^{(i)} = \sigma\left(\frac{v_t^{(i)}}{\tau}\right) \in (0,1), \tag{9}$$

where σ is the logistic sigmoid and $\tau > 0$ controls the sharpness of the transition, so that strongly evidence-dependent tokens receive more credit than marginally dependent ones instead of being weighted equally. The teacher correction is bounded in the same spirit, and the two are combined with the verifier anchor to form the modulated advantage:

$$\widetilde{A}_t^{(i)} = A_{\mathrm{ORM}}^{(i)} + \eta\, g_t^{(i)} \widetilde{v}_t^{(i)} \tanh\left(\frac{\delta_t^{(i)}}{\tau}\right), \tag{10}$$

where $\eta > 0$ controls the strength of the correction. The first term provides a trajectory-level verifier anchor, while the second redistributes credit at the token level only when the teacher correction is visually attributable. If $g_t^{(i)} = 0$, the correction vanishes and $\widetilde{A}_t^{(i)} = A_{\mathrm{ORM}}^{(i)}$, so filtering a correction falls back to outcome supervision rather than removing the token from training.

The final stage of Fig. 3 optimizes the student policy using the modulated advantage. We first define the token-level importance ratio between the current policy π_θ and the rollout policy $\pi_{\theta_{old}}$ as:

$$\rho_t^{(i)}(\theta) = \frac{\pi_\theta(\hat{y}_t^{(i)} \mid x, \hat{\mathbf{y}}_{<t}^{(i)})}{\pi_{\theta_{old}}(\hat{y}_t^{(i)} \mid x, \hat{\mathbf{y}}_{<t}^{(i)})}. \tag{11}$$

The training objective is the clipped GRPO surrogate averaged over rollouts and token positions:

$$\mathcal{L}_{\mathrm{VAG}}(\theta) = -\frac{1}{G}\sum_{i=1}^{G}\frac{1}{T_i}\sum_{t=1}^{T_i}\min\Big(\rho_t^{(i)}(\theta)\,\widetilde{A}_t^{(i)},\ \mathrm{clip}\left(\rho_t^{(i)}(\theta),\,1-\varepsilon,\,1+\varepsilon\right)\widetilde{A}_t^{(i)}\Big), \tag{12}$$

where $0 < \varepsilon < 1$ is the clipping radius. All teacher-derived signals, including δ_t, v_t, and g_t, are detached, so gradients flow to the student only through $\rho_t^{(i)}(\theta)$. Consequently, VAG introduces neither additional trainable parameters nor an auxiliary objective; it operates entirely by reshaping the advantage that GRPO already optimizes.

4 Experiment

4.1 Experimental Setup

Implementation Details. We instantiate our method using Qwen3.5-4B (Qwen Team, 2026). The model is trained on a curated synthetic dataset comprising approximately 6.2K visual-reasoning

Table 1: Comparison with single-forward-pass MLLMs and post-training methods across multi-modal reasoning benchmarks. All entries report accuracy (%).

Model	Param.	V*	Zoom	HR-Bench 4K	HR-Bench 8K	MME-RW EN	MME-RW CN	Avg.
Closed-Source Models (Single Forward Pass)								
GPT-5.2 (OpenAI, 2025)	–	79.06	50.89	81.12	78.38	72.60	68.80	71.81
GPT-5.4 (OpenAI, 2026)	–	76.96	52.66	84.00	77.88	74.20	70.93	72.77
Gemini 3.1 Pro (Google DeepMind, 2026a)	–	87.96	61.18	89.63	86.88	76.53	73.31	79.25
Gemini 3.5 Flash (Google DeepMind, 2026b)	–	89.01	61.42	89.12	86.62	75.31	73.97	79.24
Open-Source Baselines (Single Forward Pass)								
DeepEyes (Zheng et al., 2025)	7B	85.86	46.51	75.13	72.63	64.10	64.09	68.05
Thyme (Zhang et al., 2025)	7B	82.20	45.09	77.00	72.00	64.80	64.59	67.61
DeepEyesV2 (Hong et al., 2025)	7B	81.68	44.97	77.88	73.75	64.90	65.07	68.04
SenseNova-MARS (Chng et al., 2025)	8B	92.15	47.81	83.13	78.38	67.90	68.90	73.05
MiMo-VL-RL (LLM-Core-Team Xiaomi, 2025)	7B	83.25	45.68	73.50	69.38	62.73	55.89	65.07
ZwZ (Wei et al., 2026)	8B	87.96	56.69	83.63	81.75	66.57	68.09	74.12
MiniCPM-V-4.5 (Yu et al., 2025)	9B	70.68	42.60	69.63	61.50	62.65	61.64	61.45
GLM-4.6V (Zhipu AI, 2026)	106B	86.91	50.06	82.13	78.88	65.57	65.62	71.53
Qwen3-VL-Instruct (Bai et al., 2025)	235B	91.10	56.09	86.13	80.38	71.74	69.04	75.75
Qwen3.5-397B-A17B (Qwen Team, 2026)	397B	87.96	57.16	89.38	85.50	74.82	69.82	77.44
Kimi-K2.6 (Moonshot AI, 2026)	1T	88.48	53.14	81.88	78.00	69.22	66.13	72.81
Qwen3.5-4B Matched Post-Training Methods								
Qwen3.5 (Qwen Team, 2026)	4B	80.63	49.82	73.50	67.50	55.78	58.56	64.30
SFT	4B	83.25	42.60	76.50	72.12	64.10	62.50	66.85
GRPO (Shao et al., 2024)	4B	85.34	48.20	78.25	74.50	66.50	65.10	69.65
OPSD (Zhao et al., 2026)	4B	84.29	57.51	77.25	72.88	61.20	62.40	69.26
Vision-OPD (Yuan et al., 2026)	4B	89.01	65.33	82.25	81.25	74.50	70.26	77.10
VA-OPD (Liu et al., 2026)	4B	86.91	51.20	80.12	77.50	68.50	67.10	71.88
OPD-V (Aniri et al., 2026)	4B	**94.76**	65.92	85.75	83.25	76.31	74.06	80.01
VAG (Ours)	4B	94.24	**69.35**	**87.62**	**84.62**	**76.56**	**75.60**	**81.33**

examples derived from Vision-OPD (Yuan et al., 2026). During training, the privileged teacher is maintained as an EMA of the student parameters. For all experiments, the student samples $G = 8$ on-policy responses for each input. We use a training batch size of 48 and set the maximum response length to 1,024 tokens. All models undergo full-parameter training for one epoch and a 30-step warmup. All experiments are conducted using 8 NVIDIA H800 GPUs. Following the evaluation protocol of OPD-V (Aniri et al., 2026), we first perform deterministic answer matching, followed by a judge model that determines whether unresolved responses are correct by comparing them against the verified answer. For further details, see Appendices A and B.

4.2 Baselines and Benchmarks

Evaluation Benchmarks. We evaluate VAG on six visual question answering benchmarks that span fine-grained visual perception to broader multimodal problem solving. **(1) V* Bench** (Wu & Xie, 2023) tests the localization and recognition of small targets in high-resolution, visually crowded scenes. **(2) ZoomBench** (Wei et al., 2026) examines whether models can recover subtle regional evidence from a full image. **(3) HR-Bench 4K** and **(4) HR-Bench 8K** (Wang et al., 2025) evaluate fine-grained perception at native 4K and 8K resolutions, respectively. These settings are particularly challenging because conventional image resizing may discard the small-scale details required to answer the questions correctly. **(5) MME-RealWorld EN** and **(6) MME-RealWorld CN** (Zhang et al., 2024) provide the broadest setting, spanning diverse high-resolution tasks and real-world scenarios. To assess whether specializing in fine-grained perception sacrifices general multimodal ability, we further evaluate on four holdout benchmarks: **(1) MMVP** (Tong et al., 2024b), **(2) CV-Bench** (Tong et al., 2024a), **(3) MMStar** (Chen et al., 2024), and **(4) POPE** (Li et al., 2023).

Baselines. Following Vision-OPD (Yuan et al., 2026), we compare VAG against closed-source models (Google DeepMind, 2026a;b; OpenAI, 2025; 2026), agentic and large-scale open-source MLLMs (Zheng et al., 2025; Zhang et al., 2025; Hong et al., 2025; Chng et al., 2025; Bai et al., 2025; Qwen Team, 2026; Zhipu AI, 2026; Moonshot AI, 2026), and post-training baselines based on Qwen3.5-4B. The matched group includes the corresponding base models, SFT, GRPO, OPSD,

Table 2: Comparisons on holdout tasks.

Method	MMVP	CV-Bench	MMStar	POPE	Avg
Qwen3.5-4B	76.67	87.13	78.53	88.28	82.65
SFT	78.33	85.70	68.40	87.48	79.98
GRPO	79.33	87.24	70.60	86.37	80.89
DAPO	79.33	86.95	72.27	86.62	81.29
OPSD	78.68	87.27	75.07	88.83	82.46
Vision-OPD	79.67	87.27	79.60	**89.14**	83.92
OPD-V	80.00	86.07	77.40	88.38	82.96
VAG (Ours)	**84.67**	**87.91**	**79.93**	88.72	**85.31**

Table 3: Ablation on hyperparameters τ and η.

	Value	V*	Zoom	HR-Bench		MME-RW		Avg
				4K	8K	EN	CN	
	0.1	**94.24**	**69.35**	**87.62**	**84.62**	**76.56**	**75.60**	**81.33**
τ	0.2	91.62	65.68	84.12	82.75	76.91	77.52	79.77
	0.5	87.43	56.69	80.25	78.62	73.23	69.82	74.34
	0.1	**94.24**	**69.35**	**87.62**	**84.62**	**76.56**	**75.60**	**81.33**
η	0.2	86.91	58.82	79.50	77.88	70.25	69.98	73.89
	0.5	84.82	59.88	80.12	75.12	69.23	69.17	73.06

Vision-OPD, VA-OPD (Liu et al., 2026), and OPD-V. All methods in this group are evaluated at the same model scale using the same training data and evaluation protocol.

4.3 Main Results

Comparison with SOTA MLLMs. As shown in Table 1, VAG raises the Qwen3.5-4B average from 64.30% to 81.33% and exceeds OPD-V, the strongest matched baseline, by 1.32%. It attains the best result on ZoomBench, HR-Bench, MME-RealWorld, and the overall average, ranking second only on V* Bench. The margin over OPD-V is largest on ZoomBench (69.35% vs. 65.92%, +3.43%), a fine-grained benchmark that requires recovering subtle regional evidence from high-resolution images, which highlights the particular effectiveness of VAG for fine-grained perception and localization. VAG also exceeds substantially larger open-source and agentic models in average accuracy. Fig. 4 further reports results on Qwen3.5-9B, where VAG reaches 77.85% average accuracy and improves over the base model on all benchmarks.

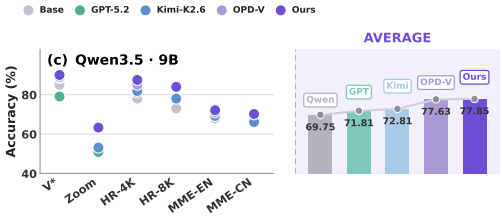


Figure 4: Performance comparison on Qwen3.5-9B.

Comparison with Post-Training Strategies. Within the matched group, VAG outperforms GRPO and OPSD by 11.68% and 12.07% on average, respectively. Notably, OPSD does not surpass GRPO (69.26% vs. 69.65%) despite providing denser supervision, indicating that unfiltered privileged corrections can offset their own benefit. Visual OPSD variants narrow this gap, yet VAG still improves over Vision-OPD by 4.23%, which demonstrates the benefit of verifying the visual attribution of token-level supervision rather than only strengthening or reweighting it.

Generalization to Holdout Tasks. Table 2 reports results on the four holdout benchmarks. VAG attains the best average accuracy (85.31%), improving over the vanilla model by 2.66% and over the strongest baseline by 1.39%. In contrast, SFT and GRPO fall below the vanilla average. VAG therefore avoids the specialization–forgetting tradeoff observed in alternative training strategies, indicating that attribution-gated supervision improves fine-grained perception without eroding general multimodal capability.

4.4 Ablations and Analysis

Impact of η and τ. Table 3 ablates the gate temperature τ and the modulation weight η. When varying τ we fix $\eta = 0.1$; when varying η we fix $\tau = 0.1$. Accuracy peaks at $\tau = 0.1$: larger temperatures flatten the gate, so visually supported and visually contradicted corrections are no longer separated. For η, 0.1 is large enough to modulate the outcome advantage yet small enough to avoid over-flipping its sign, whereas $\eta = 0.5$ degrades every benchmark. We therefore adopt $(\tau, \eta) = (0.1, 0.1)$, shown in bold, for all remaining experiments.

Modulation of Advantage. Table 4 isolates the contribution of each component of the modulated advantage under an identical training protocol. Holding all other settings fixed, we build it up from the outcome anchor A_{ORM} (Eq. 2): we successively introduce the raw distillation margin δ_t (Eq. 5), the soft visual-sensitivity $\widetilde{v}_t$ (Eq. 9), and the sign-consistent attribution gate g_t (Eq. 8) with bounded margin scaling, recovering the full form in Eq. 10. As reported in Table 4, the dense margin lifts

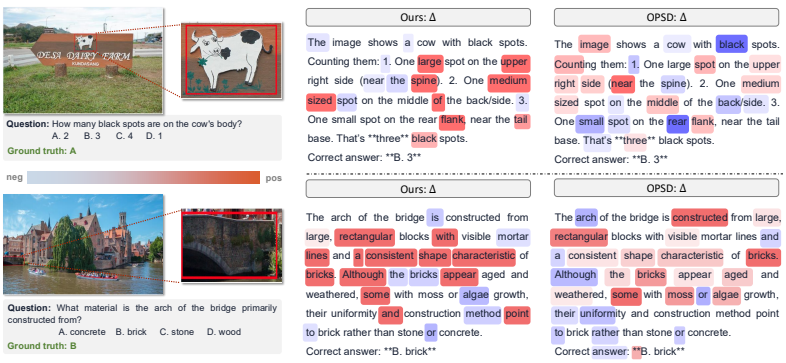


Figure 5: **Prediction and token-importance comparison.** VAG produces the correct evidence-grounded prediction and concentrates credit on decision-relevant visual tokens.

Table 4: Advantage-modulation ablation.

Correction term	V*	Zoom	HR-Bench		MME-RW		Avg
			4K	8K	EN	CN	
A_{ORM}	85.34	48.20	78.25	74.50	66.50	65.10	69.65
$+\eta\,\delta_t$	90.05	65.33	79.50	79.50	75.37	72.99	77.12
$+\eta\,\widetilde{v}_t\delta_t$	92.67	66.04	84.00	80.88	74.63	74.55	78.80
Ours	**94.24**	**69.35**	**87.62**	**84.62**	**76.56**	**75.60**	**81.33**

Table 5: Negative-view ablation.

Negative view	V*	Zoom	HR-Bench		MME-RW		Avg
			4K	8K	EN	CN	
No image	93.72	63.20	**87.88**	**87.00**	**76.82**	73.75	80.40
Black image	**94.76**	64.73	82.50	81.88	75.26	71.98	78.52
Blurred image	86.39	59.88	79.62	76.25	69.85	69.34	73.56
Random mask	94.24	**69.35**	87.62	84.62	76.56	**75.60**	**81.33**

the average from 69.65% to 77.12%, visual-sensitivity weighting further to 78.80%, and the full modulation to 81.33% ($+11.68\%$ over the outcome-only anchor), attaining the best score on every benchmark. Sensitivity weighting and sign-consistent attribution are thus complementary, jointly selecting reliable token-level corrections.

Negative Teacher View. Table 5 compares four evidence-degraded views under an identical training protocol. Random masking attains the best average accuracy (81.33%), ahead of the no-image (80.40%), black-image (78.52%), and blurred-image (73.56%) variants, indicating that v_t is most informative when the intervention removes local evidence while preserving spatial context. Its advantage concentrates on ZoomBench and MME-RealWorld CN, which depend most on resolving a small region, whereas the no-image variant remains competitive when the decisive evidence occupies a larger portion of the input.

Interpretation of VAG. Fig. 5 contrasts the token-level credit that VAG and vanilla OPSD assign to the *same* student rollout. For each token we visualize the correction term Δ_t that modulates the update, defined as $\Delta_t = \eta\, g_t^{(i)} \widetilde{v}_t^{(i)}\ \tanh(\delta_t^{(i)}/\tau)$ for VAG and as $\Delta_t = \delta_t^{(i)}$ for OPSD, with the color scale running from negative (blue) to positive (red). In the counting example (top), whose rollout is *incorrect*, OPSD spreads non-zero credit over much of the response, including function words and a strongly negative weight on "black". VAG instead concentrates credit on the tokens that *enumerate and localize* the spots ("upper", "spine", "flank"), while suppressing the erroneous count token "3". In the material example (bottom), whose rollout is *correct*, VAG weights the chain of *visual premises* supporting the answer ("mortar lines", "consistent shape"), whereas OPSD assigns comparable weight to weakly grounded tokens such as "aged" and "rather". The attribution gate thus removes sign-inconsistent teacher residuals and yields a *sparser, evidence-aligned* credit map, irrespective of whether the rollout is right or wrong.

5 Conclusion

We introduce reinforcement learning with visually attributable on-policy self-distillation to mitigate privileged-correction contamination. Our Visual-Attribution Gating anchors optimization to verifier-derived trajectory advantages and accepts token-level corrections only when consistent with the contrast between evidence-present and evidence-degraded teacher views, preserving supportive and refuting visual evidence while suppressing inconsistent residuals. Across six public benchmarks, VAG consistently outperforms state-of-the-art methods, demonstrating that visual grounding verification unifies outcome verification, token-level self-distillation, and visual attribution for multimodal policy optimization.

AI USE STATEMENT

In this work, we used generative AI tools only for language polishing of author-written text (e.g., grammar, wording, and readability) and for retrieval and discovery (e.g., finding related work). We did not use generative AI tools for research ideation, method design, code generation, experiments, data analysis, or the production of any results, figures, or other artifacts, all of which were carried out and verified by the authors. We have reviewed all AI-assisted edits to ensure that they preserve the intended technical meaning, and we take full responsibility for the final content of this work.

REFERENCES

Rishabh Agarwal, Nino Vieillard, Yongchao Zhou, Piotr Stanczyk, Sabela Ramos Garea, Matthieu Geist, and Olivier Bachem. On-policy distillation of language models: Learning from self-generated mistakes. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://arxiv.org/abs/2306.13649.

Aniri, Jinhe Bi, Peng Liao, Zengjie Jin, Volker Tresp, Fei Shen, Yunpu Ma, and Tat-Seng Chua. Opd-v: Visual on-policy self-distillation with modality balance, 2026. URL https://arxiv.org/abs/2608.05131.

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

Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Jiaqi Wang, Yu Qiao, Dahua Lin, et al. Are we on the right way for evaluating large vision-language models? *arXiv preprint arXiv:2403.20330*, 2024.

Yong Xien Chng, Tao Hu, Wenwen Tong, Xueheng Li, Jiandong Chen, Haojia Yu, Jiefan Lu, Hewei Guo, Hanming Deng, Chengjun Xie, et al. Sensenova-mars: Empowering multimodal agentic reasoning and search via reinforcement learning. *arXiv preprint arXiv:2512.24330*, 2025.

Google DeepMind. Gemini 3.1 pro. Official model release, 2026a. URL https://deepmind.google/technologies/gemini/.

Google DeepMind. Gemini 3.5 flash. Model card, 2026b. https://deepmind.google/models/model-cards/gemini-3-5-flash, accessed 2026-09-26.

Yuxian Gu, Li Dong, Furu Wei, and Minlie Huang. Minillm: On-policy distillation of large language models. *arXiv preprint arXiv:2306.08543*, 2023.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

Yinghui He, Simran Kaur, Adithya Bhaskar, Yongjin Yang, Jiarui Liu, Narutatsu Ri, Liam Fowl, Abhishek Panigrahi, Danqi Chen, and Sanjeev Arora. Self-distillation zero: Self-revision turns binary rewards into dense supervision. *arXiv preprint arXiv:2604.12002*, 2026.

Jack Hong, Chenxiao Zhao, Chenglin Zhu, Weiheng Lu, Guohai Xu, and Xing Yu. DeepEyesV2: Toward agentic multimodal model. *arXiv preprint arXiv:2511.05271*, 2025. URL https://arxiv.org/abs/2511.05271.

Jonas Hübotter, Frederike Lübeck, Lejs Behric, Anton Baumann, Marco Bagatella, Daniel Marta, Ido Hakimi, Idan Shenfeld, Thomas Kleine Buening, Carlos Guestrin, and Andreas Krause. Reinforcement learning via self-distillation. *arXiv preprint arXiv:2601.20802*, 2026. URL https://arxiv.org/abs/2601.20802.

Yiqiao Jin, Yiyang Wang, Lucheng Fu, Yijia Xiao, Yinyi Luo, Haoxin Liu, B Aditya Prakash, Josiah Hester, Jindong Wang, and Srijan Kumar. Unisd: Towards a unified self-distillation framework for large language models. *arXiv preprint arXiv:2605.06597*, 2026.

Junlong Ke, Zichen Wen, Weijia Li, Conghui He, and Linfeng Zhang. Respecting self-uncertainty in on-policy self-distillation for efficient llm reasoning. *arXiv preprint arXiv:2605.13255*, 2026.

Jeonghye Kim, Jiwon Jeon, Dongsheng Li, and Yuqing Yang. Rebellious student: Reversing teacher signals for reasoning exploration with self-distilled rlvr. *arXiv preprint arXiv:2605.10781*, 2026.

Gengsheng Li, Tianyu Yang, Junfeng Fang, Mingyang Song, Mao Zheng, Haiyun Guo, Dan Zhang, Jinqiao Wang, and Tat-Seng Chua. Unifying group-relative and self-distillation policy optimization via sample routing. *arXiv preprint arXiv:2604.02288*, 2026.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. In *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.

Wenze Lin, Jiale Zhao, Xitai Jiang, Songde Rao, Yining Li, Shenzhi Wang, Bingxiang He, and Gao Huang. On-policy distillation with verifiable reward. *arXiv preprint arXiv:2608.24696*, 2026. URL https://arxiv.org/abs/2608.24696.

Ruiqi Liu, Xiaolei Lv, Gengsheng Li, Ximo Zhu, Zhiheng Wang, Zhengbo Zhang, Junkai Chen, Zhiheng Li, Bo Li, Jun Gao, and Shu Wu. Visual-advantage on-policy distillation for vision-language models. *arXiv preprint arXiv:2605.21924*, 2026. URL https://arxiv.org/abs/2605.21924.

LLM-Core-Team Xiaomi. MiMo-VL technical report. *arXiv preprint arXiv:2506.03569*, 2025. URL https://arxiv.org/abs/2506.03569.

Moonshot AI. Kimi-K2.6. Official model release, 2026. URL https://www.moonshot.cn/.

OpenAI. GPT-5.2. Official model release, 2025. URL https://openai.com/.

OpenAI. GPT-5.4. Official model release, 2026. URL https://openai.com/.

Leyi Pan, Shuchang Tao, Yunpeng Zhai, Lingzhe Zhang, Zhaoyang Liu, Bolin Ding, Aiwei Liu, and Lijie Wen. RLCSD: Reinforcement learning with contrastive on-policy self-distillation. *arXiv preprint arXiv:2606.11709*, 2026. URL https://arxiv.org/abs/2606.11709.

Qwen Team. Qwen3.5: Towards native multimodal agents. Technical report and official model release, 2026. URL https://qwen.ai/blog?id=qwen3.5.

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. *arXiv preprint arXiv:2402.03300*, 2024. URL https://arxiv.org/abs/2402.03300.

Guobin Shen, Xiang Cheng, Chenxiao Zhao, Lei Huang, Jindong Li, Dongcheng Zhao, and Xing Yu. Anti-self-distillation for reasoning rl via pointwise mutual information. *arXiv preprint arXiv:2605.11609*, 2026.

Mingyang Song and Mao Zheng. A survey of on-policy distillation for large language models. *arXiv preprint arXiv:2604.00626*, 2026.

Haoxiang Sun, Zhihang Yi, Langxuan Deng, Yuhao Zhou, Peiqi Jia, Jian Zhao, Li Yuan, Jiancheng Lv, and Tao Wang. V-Zero: Answer-label-free on-policy distillation with contrastive evidence gating for fine-grained visual reasoning. *arXiv preprint arXiv:2606.25319*, 2026. URL https://arxiv.org/abs/2606.25319.

Shengbang Tong, Ellis Brown, Penghao Wu, Sanghyun Woo, Manoj Middepogu, Sai Charitha Akula, Jihan Yang, Shusheng Yang, Adithya Iyer, Xichen Pan, Austin Wang, Rob Fergus, Yann LeCun, and Saining Xie. Cambrian-1: A fully open, vision-centric exploration of multimodal LLMs. *arXiv preprint arXiv:2406.16860*, 2024a.

Shengbang Tong, Zhuang Liu, Yuexiang Zhai, Yi Ma, Yann LeCun, and Saining Xie. Eyes wide shut? exploring the visual shortcomings of multimodal LLMs. *arXiv preprint arXiv:2401.06209*, 2024b.

Jiaxuan Wang, Xuan Ouyang, Zhiyu Chen, Yulan Hu, Zheng Pan, Xin Li, and Lan-Zhe Guo. Trace: Distilling where it matters via token-routed self on-policy alignment. *arXiv preprint arXiv:2605.10194*, 2026.

Wenbin Wang, Liang Ding, Minyan Zeng, Xiabin Zhou, Li Shen, Yong Luo, Wei Yu, and Dacheng Tao. Divide, conquer and combine: A training-free framework for high-resolution image perception in multimodal large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 7907–7915, 2025.

Lai Wei, Liangbo He, Jun Lan, Lingzhong Dong, Yutong Cai, Siyuan Li, Huijia Zhu, Weiqiang Wang, Linghe Kong, Yue Wang, Zhuosheng Zhang, and Weiran Huang. Zooming without zooming: Region-to-image distillation for fine-grained multimodal perception. *arXiv preprint arXiv:2602.11858*, 2026.

Penghao Wu and Saining Xie. V*: Guided visual search as a core mechanism in multimodal LLMs. *arXiv preprint arXiv:2312.14135*, 2023.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

Chenxu Yang, Chuanyu Qin, Qingyi Si, Minghui Chen, Naibin Gu, Dingyu Yao, Zheng Lin, Weiping Wang, Jiaqi Wang, and Nan Duan. Self-distilled RLVR. *arXiv preprint arXiv:2604.03128*, 2026. URL https://arxiv.org/abs/2604.03128.

Hee Suk Yoon, Eunseop Yoon, Jaehyun Jang, SooHwan Eom, Ji Woo Hong, Mark Hasegawa-Johnson, Qi Dai, Chong Luo, and Chang D. Yoo. Decomposed on-policy distillation for vision-language reasoning: Steering gradients for visual grounding, 2026. URL https://arxiv.org/abs/2606.00564.

Tianyu Yu, Zefan Wang, Chongyi Wang, Fuwei Huang, Wenshuo Ma, Zhihui He, Tianchi Cai, Weize Chen, Yuxiang Huang, Yuanqian Zhao, et al. MiniCPM-V 4.5: Cooking efficient MLLMs via architecture, data, and training recipe. *arXiv preprint arXiv:2509.18154*, 2025. URL https://arxiv.org/abs/2509.18154.

Qianhao Yuan, Jie Lou, Xing Yu, Hongyu Lin, Le Sun, Xianpei Han, and Yaojie Lu. Vision-OPD: Learning to see fine details for multimodal LLMs via on-policy self-distillation. *arXiv preprint arXiv:2605.18740*, 2026. URL https://arxiv.org/abs/2605.18740.

Kangning Zhang, Yixing Li, Shuai Shao, Qingyao Li, Zhengxi Lu, Zhiyuan Yao, Jianghao Lin, Wenxiang Jiao, Yuan Lu, Weiwen Liu, et al. Vad: Attributing visual evidence for target reconstruction in multimodal on-policy distillation. *arXiv preprint arXiv:2607.28590*, 2026.

Yi-Fan Zhang, Huanyu Zhang, Haochen Tian, Chaoyou Fu, Shuangqing Zhang, Junfei Wu, Feng Li, Kun Wang, Qingsong Wen, Zhang Zhang, Liang Wang, Rong Jin, and Tieniu Tan. MME-RealWorld: Could your multimodal LLM challenge high-resolution real-world scenarios that are difficult for humans? *arXiv preprint arXiv:2408.13257*, 2024.

Yi-Fan Zhang, Xingyu Lu, Shukang Yin, Chaoyou Fu, Wei Chen, Xiao Hu, Bin Wen, Kaiyu Jiang, Changyi Liu, Tianke Zhang, and Guorui Zhou. Thyme: Think beyond images. *arXiv preprint arXiv:2508.11630*, 2025. URL https://arxiv.org/abs/2508.11630.

Siyan Zhao, Zhihui Xie, Mengchen Liu, Jing Huang, Guan Pang, Feiyu Chen, and Aditya Grover. Self-distilled reasoner: On-policy self-distillation for large language models. *arXiv preprint arXiv:2601.18734*, 2026. URL https://arxiv.org/abs/2601.18734.

Ziwei Zheng, Michael Yang, Jack Hong, Chenxiao Zhao, Guohai Xu, Le Yang, Chao Shen, and Xing Yu. DeepEyes: Incentivizing "thinking with images" via reinforcement learning. *arXiv preprint arXiv:2505.14362*, 2025. URL https://arxiv.org/abs/2505.14362.

Zhipu AI. GLM-4.6V. Official model release, 2026. URL https://www.zhipuai.cn/.

A Evaluation Protocol for Benchmarks

To ensure that every method is compared under identical conditions, we adopt a single evaluation protocol and apply it uniformly to VAG and to all reproduced baselines. Table 6 lists the number of evaluation samples and the scoring rule used for each benchmark.

Following OPD-V (Aniri et al., 2026), scoring proceeds in two stages. We first apply deterministic answer matching, which resolves the majority of responses without any model in the loop. For the multiple-choice benchmarks, a response counts as correct if it matches either the answer text or the option letter, after normalizing surface variations such as capitalization, surrounding punctuation, and common answer prefixes. Only the responses that remain unresolved are passed to the second stage, where a Qwen3.5-series (Qwen Team, 2026) model acts as a binary Yes/No judge and decides whether the response is consistent with the verified answer.

Both stages are necessary for the benchmarks we evaluate. Responses on fine-grained perception tasks frequently state the correct answer inside a longer justification, which deterministic matching alone would count as wrong; delegating every response to the judge instead would introduce scoring variance unrelated to the method under evaluation. Calling the judge only on the unresolved responses therefore keeps most decisions exact while still absorbing benign paraphrasing. Since the same protocol, judge model, and prompts apply to every entry in Table 1, the reported differences reflect the trained policies rather than the evaluation.

B Training Prompt Example

Fig. 6 presents one complete training instance: the original image observed by the student, the two teacher images derived from it, and the textual prompt used in each pass.

The three image conditions differ only in the visual evidence they expose. The student receives the original, unmodified image and must therefore locate the relevant detail on its own. Both teacher images are constructed from this same image. The positive teacher receives the evidence-centered crop defined by the annotated bounding box, so the decisive detail occupies a far larger fraction of the visual field. The negative teacher receives the same crop after random masking, which weakens the local evidence while preserving the surrounding spatial context.

The textual prompt is assembled from a fixed set of fields: an image placeholder, the question, a visual focus instruction that directs the model to the annotated region, the answer choices, and an output-format instruction that requests only the option letter. The ground-truth answer is consumed by the verifier and is never shown to the model. Constraining the output format yields two additional benefits. First, it reduces the fraction of responses that require the Yes/No judge at evaluation time. Second, it leaves little room for stylistic variation between the teacher and student branches, so the gap δ_t is less likely to reflect differences in wording instead of differences in visual evidence.

Because the question, answer choices, output-format instruction, and the student-generated prefix are identical across the three passes, the only controlled difference between the two teacher branches is the availability of the target evidence. This is what allows the visual contrast v_t in Eq. 7 to be interpreted as a statement about visual evidence rather than about prompt or sampling differences. Both teacher views are used only during training; at inference time the deployed policy observes the original image alone and requires no privileged input.

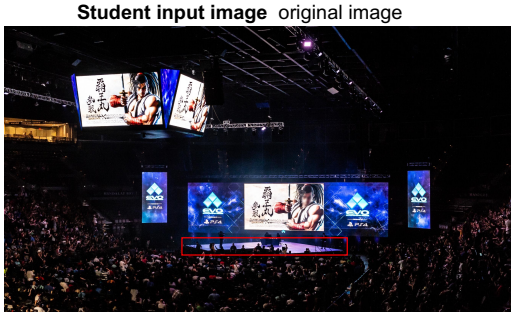
C VAG Algorithm

Algorithm 1 states the full VAG training loop and makes explicit the order in which the outcome signal and the token-level signal are combined.

Each iteration begins with rollout collection. For each input $x = (q, I)$, the student samples G on-policy responses $\hat{\mathbf{y}}^{(i)}$. The verifier assigns each response a reward $R^{(i)}$, which is normalized within the group to obtain $A^{(i)}_{\mathrm{ORM}}$ as in Eq. 2. This advantage is shared by all tokens of the response. The two teacher branches then replay each sampled response using the same student-generated prefix $\hat{\mathbf{y}}^{(i)}_{<t}$. The sampled-token probabilities $p^{(i)}_S(\hat{y}_t)$, $p^{(i)}_{T+}(\hat{y}_t)$, and $p^{(i)}_{T-}(\hat{y}_t)$ are defined in Eqs. 3, 4, and 6, respectively.

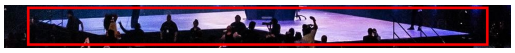
Table 6: Number of evaluation samples and scoring rule for each benchmark. All methods in Table 1 are evaluated with this identical protocol.

Benchmark	Samples	Scoring Rule
V* Bench	191	Rule-based answer matching is applied first; unresolved responses are compared with the verified answer by a Yes/No judge.
ZoomBench	845	Generic answer matching handles normalized responses first; unresolved cases use the verified-answer Yes/No judge.
HR-Bench 4K	800	Multiple-choice matching checks direct answer matches and option-letter matches before the Yes/No judge handles unresolved responses.
HR-Bench 8K	800	Multiple-choice matching follows the HR-Bench 4K rule.
MME-RealWorld EN	23609	Multiple-choice matching and the Yes/No judge are applied to the English subset.
MME-RealWorld CN	5917	The Chinese subset uses the MME-RealWorld multiple-choice scoring rule.



Student input image original image

Positive teacher input image
Zoom-in-image cropped image from the red bounding box



Negative teacher input image
Masked Image constructed by randomly masking the Zoom-in-image



Prompt image: Student input image.

Image placeholder: <image>

Question: What color is the lighting on the stage floor?

Visual focus instruction: Only focus on the objects inside the red bounding box in the image to answer this question.

Answer choices: A. black B. white C. blue D. purple

Output format: Answer with the option's letter from the given choices.

Ground-truth answer: D

Figure 6: **Training prompt and paired visual inputs.** The textual prompt, comprising the image placeholder, question, visual focus instruction, answer choices, and output format, is shared across all three passes. The student is conditioned on the original image, the positive teacher on the evidence-centered crop given by the red bounding box, and the negative teacher on a randomly masked version of that crop. The ground-truth answer is used only by the verifier.

From these probabilities, the algorithm computes the dense teacher correction $\delta_t^{(i)}$ and visual evidence contrast $v_t^{(i)}$ using the stop-gradient log-ratios in Eqs. 5 and 7. The attribution gate $g_t^{(i)}$ retains corrections only when $\delta_t^{(i)} v_t^{(i)} > 0$ (Eq. 8). The gate checks whether the correction and visual contrast have the same sign, rather than comparing their magnitudes. Thus, only corrections whose direction agrees with the visual evidence are retained. Admitted corrections are bounded by $\tanh(\delta_t^{(i)}/\tau)$ and weighted by $\eta\, \widetilde{v}_t^{(i)}$, then added to $A_{\mathrm{ORM}}^{(i)}$ to obtain $\widetilde{A}_t^{(i)}$ as in Eq. 10.

Two properties of the loop are worth noting. First, all teacher-derived quantities are detached, so gradients reach the student only through $\rho_t^{(i)}(\theta)$ in $\mathcal{L}_{\mathrm{VAG}}(\theta)$; the teacher acts as a signal source rather than as a differentiable path. Second, when $g_t^{(i)} = 0$, the modulated advantage satisfies $\widetilde{A}_t^{(i)} = A_{\mathrm{ORM}}^{(i)}$. Thus, rejecting a correction preserves the token's verifier-derived supervision. After each student update, the teacher parameters are updated by EMA to track the student policy without separate teacher training.

Algorithm 1 VAG: Visual-Attribution Gated RLVR

Require: Training set $\mathcal{D}$, student policy π_θ, EMA teacher π_ϕ, verifier $\mathcal{V}$, group size G, EMA rate α, sharpness τ, correction scale η
Ensure: Updated policy parameters θ

Initialize teacher parameters $\phi \leftarrow \theta$
for each training step **do**
  Sample a mini-batch $\mathcal{B} \subset \mathcal{D}$
  for each $(x, a^\star) \in \mathcal{B}$ with $x = (q, I)$ **do** ▷ $a^\star$ is the ground-truth answer for x
    Obtain the evidence-preserving view I_+ from the training annotation and set $x_+ = (q, I_+)$
    $I_- \leftarrow \mathrm{Mask}(I_+)$ and $x_- = (q, I_-)$ ▷ evidence-degraded negative view
    for $i = 1, \ldots, G$ **do**
      Sample rollout $\hat{\mathbf{y}}^{(i)} = (\hat{y}_1^{(i)}, \ldots, \hat{y}_{T_i}^{(i)}) \sim \pi_\theta(\cdot \mid x)$
      $R^{(i)} \leftarrow \mathcal{V}(\hat{\mathbf{y}}^{(i)}, a^\star)$ ▷ verifiable outcome reward
    end for
    $A_{\mathrm{ORM}}^{(i)} \leftarrow \dfrac{R^{(i)} - \mu_G}{\sigma_G}$ for all i ▷ group-relative outcome advantage, shared by all tokens
    for $i = 1, \ldots, G$ **do**
      for $t = 1, \ldots, T_i$ **do**
        $p_S^{(i)}(\hat{y}_t) \leftarrow \pi_\theta(\hat{y}_t^{(i)} \mid x, \hat{\mathbf{y}}_{<t}^{(i)})$
        $p_{T+}^{(i)}(\hat{y}_t) \leftarrow \pi_\phi(\hat{y}_t^{(i)} \mid x_+, \hat{\mathbf{y}}_{<t}^{(i)})$ ▷ positive teacher
        $p_{T-}^{(i)}(\hat{y}_t) \leftarrow \pi_\phi(\hat{y}_t^{(i)} \mid x_-, \hat{\mathbf{y}}_{<t}^{(i)})$ ▷ negative teacher
        $\delta_t^{(i)} \leftarrow \mathrm{sg}\Big[\log p_{T+}^{(i)}(\hat{y}_t) - \log p_S^{(i)}(\hat{y}_t)\Big]$ ▷ dense teacher correction, Eq. 5
        $v_t^{(i)} \leftarrow \mathrm{sg}\Big[\log p_{T+}^{(i)}(\hat{y}_t) - \log p_{T-}^{(i)}(\hat{y}_t)\Big]$ ▷ visual evidence contrast, Eq. 7
        $g_t^{(i)} \leftarrow \mathbb{1}\Big[\delta_t^{(i)}\, v_t^{(i)} > 0\Big]$ ▷ visual attribution gate, Eq. 8
        $\widetilde{v}_t^{(i)} \leftarrow \sigma\Big(v_t^{(i)}/\tau\Big)$ ▷ bounded visual-sensitivity weight, Eq. 9
        $\widetilde{A}_t^{(i)} \leftarrow A_{\mathrm{ORM}}^{(i)} + \eta\, g_t^{(i)}\, \widetilde{v}_t^{(i)}\, \tanh\Big(\delta_t^{(i)}/\tau\Big)$ ▷ verifier-anchored, visually attributed advantage, Eq. 10
      end for
    end for
  end for
  Update θ by minimizing $\mathcal{L}_{\mathrm{VAG}}$ (Eq. 12) with advantages $\{\widetilde{A}_t^{(i)}\}$ over $\mathcal{B}$
  $\phi \leftarrow (1 - \alpha)\, \phi + \alpha\, \theta$
end for

D Training Dynamics

Fig. 7 reports two statistics throughout training: the mean verifier reward within each rollout group (left) and the mean response length (right). The reward curve shows a clear upward trend, increasing from approximately 0.4–0.5 over the first twenty steps to 0.65–0.75 in the second half of training, before gradually stabilizing. This trend suggests that our training strategy consistently improves the policy while maintaining stable optimization in the later stages of training.

The response-length curve exhibits a different pattern. It oscillates within a relatively stable range throughout training and ends at a level comparable to its initial value, with no sustained growth. This is important because response-length inflation is a common confound in reinforcement learning with verifiable rewards: a policy may increase its pass rate by producing longer responses that enumerate more candidate answers. In such cases, an increasing reward curve may reflect a change in response strategy rather than genuine improvement in visual perception.

Taken together, these two curves support the interpretation of our quantitative results. Gating corrections by visual attribution sharpens token-level credit assignment without destabilizing the policy's generation behavior, consistent with retaining A_{ORM} as the optimization anchor rather than replacing it with dense teacher supervision.

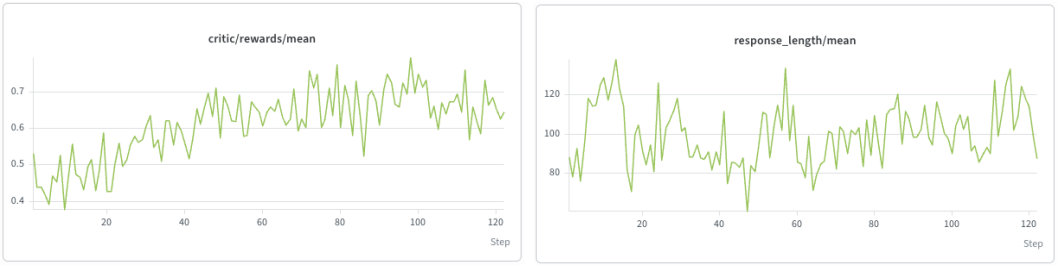


Figure 7: **Training dynamics of VAG.** Left: the mean verifier reward increases over training. Right: the mean response length fluctuates within a broadly constant band and shows no sustained growth. Both panels report raw per-step training statistics without smoothing.

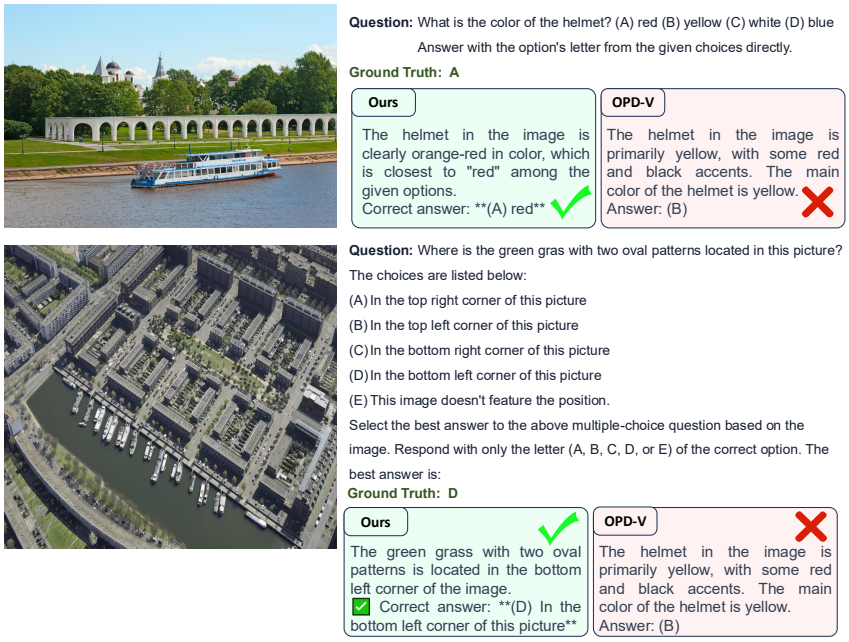


Figure 8: **Fine-grained attribute recognition and spatial localization.** Top: VAG reads the small helmet region as orange-red and selects the closest option, red, matching the ground truth, whereas OPD-V reports yellow. Bottom: VAG correctly localizes the green region with two oval patterns in the bottom-left corner. OPD-V answers incorrectly in both cases.

E Additional Qualitative Comparisons

Figs. 8, 9, and 10 present qualitative comparisons between VAG and OPD-V (Aniri et al., 2026) on the test set. Fig. 8 covers fine-grained attribute recognition and spatial localization. In the attribute case, the question asks for the color of a helmet that occupies a small portion of the frame. VAG describes the region as orange-red and maps it to the closest available option, red, matching the ground truth, whereas OPD-V reports the helmet as predominantly yellow. The two answers differ in what the model reads off the region rather than in how it reasons afterwards, which is precisely the failure mode that dense but unattributed corrections are least able to repair. In the localization case, the question asks where a green region with two oval patterns appears, and VAG places it in the bottom-left corner in agreement with the annotation.

Fig. 9 shows an object-part recognition case in which the model must name the car part visible on the right side of the image. VAG enumerates the options against the visible content, rules out the headlight and the side mirror as absent from the frame, and selects the taillight, which matches the ground truth. Fig. 10 is the most demanding case, because the visual reading must survive several arithmetic steps. The question asks for the average of two bars in a thirty-bar chart whose x-axis is labeled only every fifth item. VAG initially misaligns the labels with the bars and obtains a value that is absent from the answer choices; it then re-examines the axis, re-derives the bar-to-item mapping,

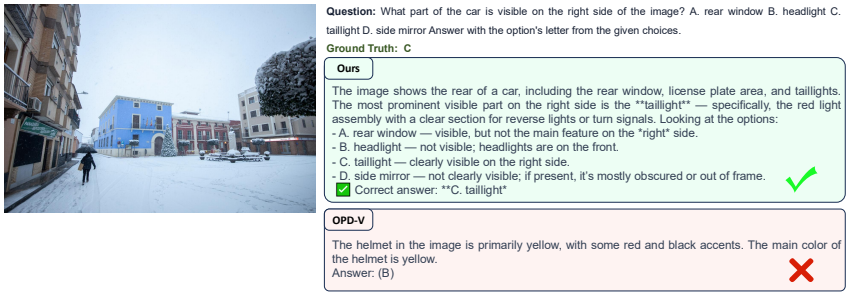


Figure 9: **Object-part recognition.** VAG rules out the parts that are not visible in the frame and identifies the part on the right side of the car as the taillight (option C), matching the ground truth, whereas OPD-V selects an incorrect option.

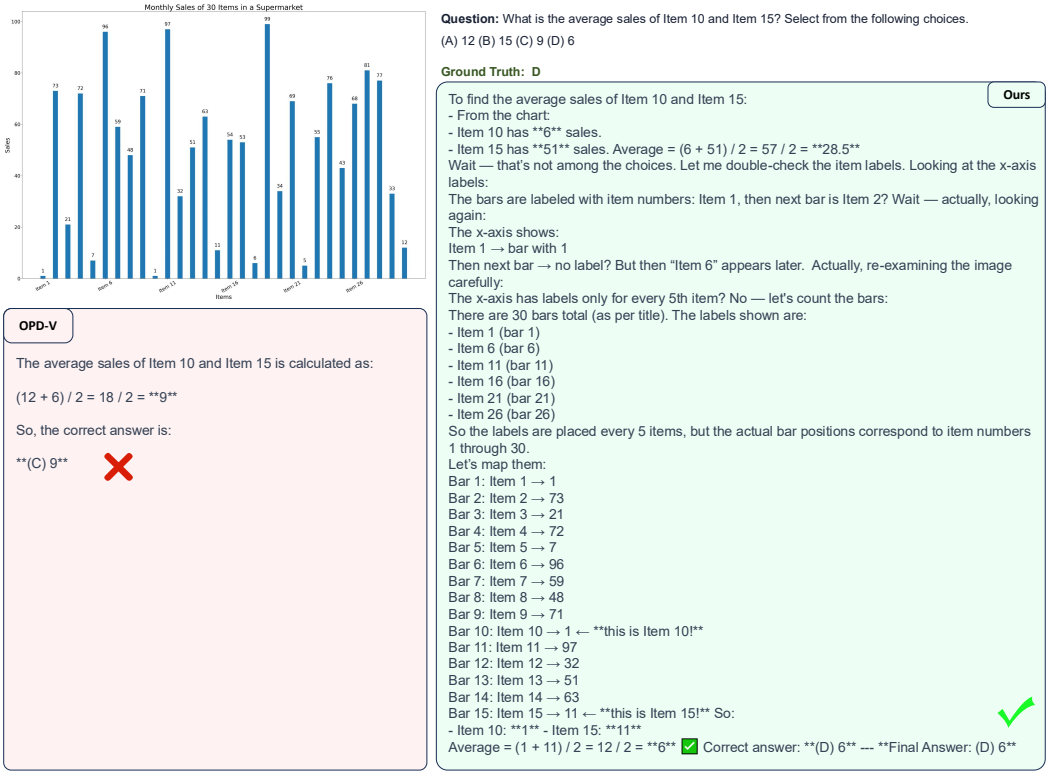


Figure 10: **Chart-based visual arithmetic.** The x-axis is labeled only every fifth item. VAG detects that its initial reading yields a value absent from the answer choices, re-derives the bar-to-item mapping, reads Item 10 and Item 15 as 1 and 11, and returns their average 6 (option D). OPD-V uses unsupported bar values and returns 9 (option C).

reads the two bars as 1 and 11, and computes their average as 6, which is correct. OPD-V instead commits to unsupported bar values and reports 9. The example shows that the benefit of attribution is not confined to perception: an incorrectly read value propagates into every subsequent step, so reading the chart correctly is a precondition for the arithmetic to be meaningful at all.

Across these cases the pattern is consistent with the aggregate results in Table 1. The largest gains appear where the answer hinges on a localized detail that must be read correctly before any further reasoning can help.