SEEING WHAT MATTERS: REWARD-ANCHORED VISUAL CREDIT ASSIGNMENT VIA SELF-DISTILLED REINFORCEMENT LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Reinforcement learning with verifiable rewards improves multimodal reasoning through outcome-level feedback, yet its sequence-level advantage is broadcast across all generated tokens. This obscures which tokens are associated with decisive visual evidence during generation and limits reliable learning from visual evidence. A natural way to address this limitation is to introduce privileged visual evidence and contrast model predictions across different visual conditions, thereby providing finer-grained learning signals for vision-critical tokens. However, the token-level differences induced by these visual conditions do not in themselves constitute reliable credit signals. They may be confounded by factors unrelated to fine-grained visual evidence and may not align with the optimization direction indicated by outcome-level rewards. To this end, we propose **VisCA**, a self-distilled reinforcement learning framework for reward-anchored visual credit assignment. VisCA treats privileged visual differences as candidate evidence for token-level credit assignment rather than directly using them as credit signals. It introduces a matched counterfactual visual condition to assess whether these differences are attributable to fine-grained visual evidence. The verified signal is then calibrated across sibling rollouts. It is subsequently used to modulate the verifier-derived outcome advantage at visually informative tokens, without changing the reward-determined optimization direction. Across six fine-grained visual understanding benchmarks, VisCA improves the average score over the corresponding base models by 7.04 points at 4B scale and 6.21 points at 9B scale. It also maintains strong average performance on held-out tasks that probe broader multimodal capabilities.

1 INTRODUCTION

Multimodal Large Language Models (MLLMs) have achieved strong performance in general visual understanding and reasoning (Google DeepMind, 2025; OpenAI, 2025). However, they remain vulnerable on fine-grained visual understanding tasks, where correct answers often hinge on small but decisive evidence embedded in complex images (Liu et al., 2025; Khayatkhoei et al., 2025; Wang et al., 2026a). A model may overlook a number, confuse an attribute, or misread a spatial relation, while still producing a fluent and seemingly plausible response grounded in misinterpreted visual evidence (Zhang et al., 2026). Such failures pose a distinctive challenge in identifying and effectively learning from the visual evidence that truly supports the correct answer.

Among current approaches to post-training MLLMs, reinforcement learning with verifiable rewards (RLVR) has emerged as a major paradigm, using outcome-level feedback to improve model reasoning capabilities (Guo et al., 2025; Yang et al., 2025; Hong et al., 2025). In widely used group-based methods such as GRPO (Shao et al., 2024), each sampled response receives a sequence-level advantage derived from its final reward, and the same advantage is applied across all generated tokens. As illustrated in Figure 1(a), such coarse-grained credit assignment determines whether a response should be reinforced or penalized. However, it provides little guidance on which tokens are tied to decisive visual evidence. This exposes a central challenge for multimodal RLVR in assigning fine-grained visual credit while preserving the optimization direction determined by outcome rewards.

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.

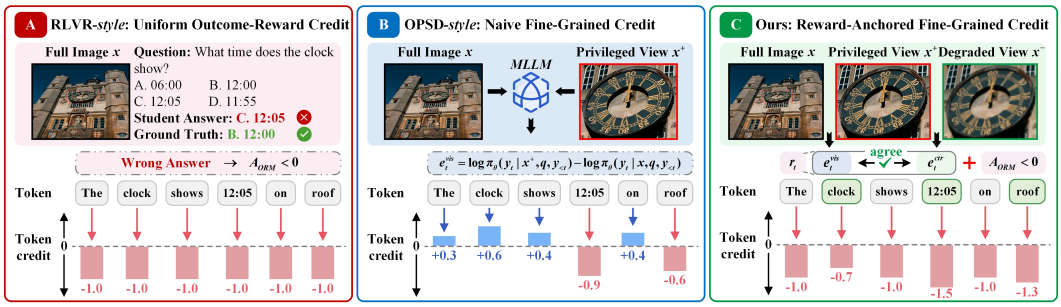

Figure 1: Comparison of credit assignment strategies: **(a)** uniform credit from outcome rewards, **(b)** fine-grained credit from privileged differences, and **(c)** our reward-anchored fine-grained credit.

Recent on-policy self-distillation (OPSD) methods (Zhao et al., 2026; Hübotter et al., 2026; Zhang et al., 2026; Bi et al., 2026) provide a complementary source of fine-grained supervision for RLVR by using privileged information to derive token-level supervision along student-sampled trajectories. In fine-grained visual understanding, a natural form of such privilege is an evidence-centered regional view (Wei et al., 2026; Yuan et al., 2026). The same MLLM can often resolve the answer when the relevant region is isolated even when it fails on the full image. This suggests that changes in its predictions across these visual conditions contain useful information about where visual evidence influences generation. However, directly using the discrepancy between privileged-view and full-image predictions for visual credit is problematic. As illustrated in Figure 1(b), privileged-view corrections can be source-mixed. Strong disagreement does not necessarily mean that the corresponding token-level change is explained by the decisive visual evidence itself. This ambiguity limits the direct use of privileged visual differences for fine-grained credit assignment.

These observations motivate a different formulation of privileged visual supervision. Privileged visual differences should be treated as candidate evidence for credit assignment rather than as credit signals themselves. Reliable visual credit assignment must address two distinct questions. *First, is a token-level difference genuinely attributable to fine-grained visual evidence? Second, once a token is identified as visually relevant, in which direction should the policy be updated?* We argue that these questions should be handled by different signals. Controlled interventions on visual evidence can assess whether token-level differences are attributable to the relevant visual evidence, whereas verifier-derived rewards should determine the direction of the policy update.

Based on this principle, we propose **VisCA**, a self-distilled reinforcement learning framework for reward-anchored visual credit assignment. As illustrated in Figure 1(c), VisCA first treats privileged visual differences as candidate token-level evidence rather than supervision signals. It then verifies whether these differences are attributable to fine-grained visual evidence through a matched counterfactual view. The resulting visual evidence is further calibrated across sibling rollouts to emphasize signals that consistently contribute to the sampled responses. Rather than replacing the verifier-derived advantage, VisCA uses this evidence only to modulate the magnitude of token-level updates while preserving the reward-determined update direction. Drawing inspiration from prior work (Pan et al., 2026), a dual-path optimization objective further prevents the sparse visually modulated signal from being diluted by the remaining tokens. In this way, VisCA converts privileged visual differences into fine-grained credit that is both visually grounded and reward anchored.

We evaluate VisCA on six fine-grained visual understanding benchmarks with 4B and 9B models. Across both scales, VisCA consistently outperforms GRPO and representative self-distillation baselines. It improves the six-benchmark average over the corresponding base model by 7.04 points at 4B and 6.21 points at 9B. VisCA also maintains strong performance on held-out multimodal evaluations, indicating that these gains do not come at the expense of broader multimodal capabilities.

Our contributions are summarized as follows:

- We identify a key ambiguity in using privileged visual differences for fine-grained credit assignment. Such differences indicate where predictions change, but do not guarantee visual attribution or alignment with reward-determined updates.
- We propose VisCA, which verifies token-level visual evidence through matched counterfactual views and uses the verified evidence to refine verifier-derived credit while preserving the reward-determined update direction.

- We demonstrate consistent gains across six fine-grained visual benchmarks at both 4B and 9B scales while maintaining strong held-out multimodal performance.

2 Related Work

2.1 Credit Assignment in RLVR

Reinforcement learning with verifiable rewards (RLVR), exemplified by GRPO (Shao et al., 2024), has become a standard paradigm for post-training reasoning models. However, its sequence-level advantage is typically applied uniformly across generated tokens, motivating finer-grained credit assignment. Existing approaches either introduce process rewards or step-level value estimators (Luo et al., 2024; Cui et al., 2025; Lightman et al., 2024; Wang et al., 2024), or reshape token-level credit using signals such as uncertainty, entropy, attention, and outcome sensitivity (Xie et al., 2026; Wang et al., 2026b; Li et al., 2025; 2026b). While these methods improve credit granularity, they often require auxiliary models or rely on indirect proxies for token importance. More recent self-distillation methods provide dense token-level signals within RLVR (Heakl et al., 2026). RLSD (Yang et al., 2026) modulates verifier-derived updates using privileged teacher–student discrepancies, while RLCSD (Pan et al., 2026) further contrasts privileged contexts. However, in fine-grained visual understanding, privileged visual differences indicate where predictions change but not whether those changes are attributable to decisive visual evidence. VisCA therefore treats them as candidate evidence and uses a matched counterfactual view to verify whether they are attributable to fine-grained visual evidence before using them to refine verifier-derived credit.

2.2 On-policy Self-Distillation

On-policy distillation (OPD) (Agarwal et al., 2024) provides dense token-level supervision along student-sampled trajectories, but typically relies on a separate teacher throughout training. On-policy self-distillation (OPSD) (Zhao et al., 2026; Cai et al., 2026) removes this dependency by using the same model as both teacher and student, with the teacher conditioned on privileged information such as reference solutions (Zhao et al., 2026) or environmental feedback (Hübotter et al., 2026). In fine-grained visual understanding, privileged information can naturally take the form of enhanced visual evidence. This has motivated methods that construct privileged supervision by manipulating the model's visual context (Yuan et al., 2026; Li et al., 2026a). Vision-OPD (Yuan et al., 2026) distills perception from an evidence-centered crop, OPD-V (Bi et al., 2026) uses paired zoomed-in and masked views to regulate selective self-distillation, and VAD (Zhang et al., 2026) contrasts evidence-present and evidence-removed views to construct visually attributed targets. These methods use privileged visual conditions to construct or regulate distillation supervision. VisCA instead treats visual contrasts as diagnostic evidence for attribution and uses the verified signal only to refine verifier-derived credit, leaving the optimization direction determined by outcome rewards.

3 Methodology

3.1 Preliminary

We define token-level visual dependency as the rectified log-probability gain of a sampled token under the privileged view over the full image. It reflects reliance on fine-grained visual information.

Definition 3.1 (**Token-level visual dependency**). *Let x denote the original full image and x^+ denote the corresponding privileged view. At a generation step t with prefix $y_{<t}$, the token-level visual dependency is defined as:*

$$VD_t = max(\log p_T\big(y_t \mid x^+, q, y_{<t}\big) - \log p_T(y_t \mid x, q, y_{<t})\,, 0), \tag{1}$$

A high VD_t suggests that predicting y_t relies more strongly on the fine-grained visual information exposed by the privileged view.

Observation 1: Visual Dependency is Sparse at the Token-level. With the above metric, we first examine how visual dependency is distributed across token positions within individual rollouts. Figure 2(a) shows a highly skewed distribution toward small values. Specifically, 87.1% of tokens have

visual dependency below 0.05, whereas only 1.07% exceed 0.5. Fine-grained visual information therefore does not influence all generation steps uniformly. Instead, its effect is concentrated on a sparse subset of token positions. Treating all tokens equally can consequently dilute the learning signal associated with visually informative decisions.

Observation 2: Visual Dependency is Heterogeneous across Rollouts. We further aggregate token-level visual dependency to the rollout level by averaging the dependency scores over the full rollout. The resulting score reflects the rollout's overall reliance on fine-grained visual information. As shown in Figure 2(b), rollout-level visual dependency varies substantially across sampled responses. Most rollouts exhibit weak visual dependency (84.9% below 0.05), with a pronounced high-dependency tail reaching 0.109/0.336 at the 95th/99th percentiles. This indicates that different rollouts can rely on fine-grained visual information to markedly different degrees. Treating their visual signals uniformly therefore overlooks meaningful variation in visual reliance across rollouts.

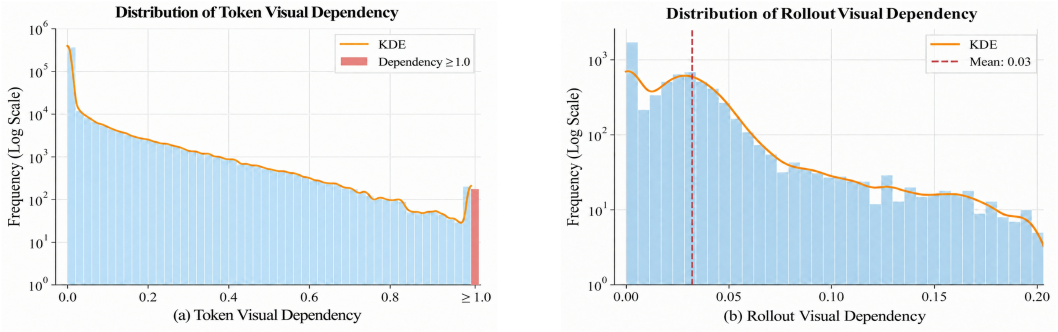

Figure 2: Distribution of visual dependency at the token and rollout levels.

Together, these observations reveal a two-level non-uniformity in visual dependency. Within each rollout, visual dependency is concentrated on a sparse subset of tokens, while its overall strength varies substantially across rollouts. This motivates visual credit assignment to account for both levels by selectively identifying informative token-level signals and calibrating their contributions according to differences across sibling rollouts.

3.2 The Proposed VisCA

We propose **VisCA**, a self-distilled reinforcement learning framework for fine-grained visual credit assignment (Figure 3). VisCA verifies token-level visual evidence with matched counterfactuals (Section 3.3), calibrates it across sibling rollouts (Section 3.4), and modulates verifier-derived advantages without changing the reward-determined direction (Section 3.5).

3.3 Token-level Visual Credit Extraction

For an image-question pair (x, q), we consider an evidence-centered region R containing the visual information required to answer the question. The policy π_θ observes the full image and samples an on-policy response $y = (y_1, \ldots, y_T)$. We construct a privileged regional view x^+ from R and a degraded counterpart x^- that preserves the same region while suppressing its fine-grained details. For each student-generated prefix $y_{<t}$, an EMA self-teacher $\pi_{\bar{\theta}}$, maintained as a moving average of the policy, evaluates the likelihood of the sampled token y_t under these visual conditions.

Privileged Visual Evidence Construction. The privileged regional view x^+ provides a focused view of the task-relevant region compared with the full-image observation x. We quantify the token-level change induced by this privileged visual condition using the log-likelihood difference:

$$e_t^{\text{vis}} = \log \pi_{\bar{\theta}} \left(y_t \mid x^+, q, y_{<t} \right) - \log \pi_{\bar{\theta}} \left(y_t \mid x, q, y_{<t} \right). \tag{2}$$

A positive value of e_t^{vis} indicates that the privileged visual condition increases the likelihood of the sampled token y_t, suggesting that the prediction is sensitive to the exposed fine-grained visual information. We treat this privileged visual difference as candidate token-level visual evidence.

Counterfactual Reliability Verification. The privileged visual difference in Eq. (2) may capture not only fine-grained visual evidence but also incidental changes introduced by the regional trans-

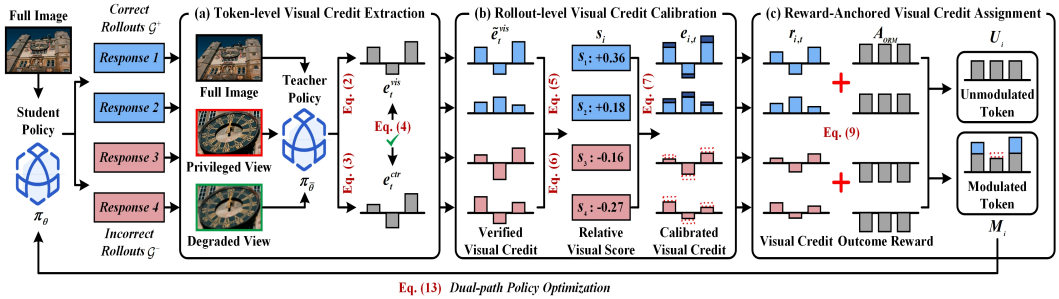

Figure 3: Overview of the proposed VisCA framework.

formation, such as altered context or visual composition. To verify whether the observed token-level change is consistent with the presence of relevant visual details, we compare the evidence-present view x^+ with the matched evidence-suppressed view x^-:

$$e_t^{\mathrm{ctr}} = \log \pi_{\bar{\theta}}\left(y_t \mid x^+, q, y_{<t}\right) - \log \pi_{\bar{\theta}}\left(y_t \mid x^-, q, y_{<t}\right). \tag{3}$$

We measure counterfactual consistency by the directional agreement between e_t^{vis} and e_t^{ctr}:

$$g_t = \mathbb{I}\left(e_t^{\mathrm{vis}} e_t^{\mathrm{ctr}} > 0\right), \qquad \tilde{e}_t^{\mathrm{vis}} = g_t \cdot e_t^{\mathrm{vis}}. \tag{4}$$

The matched counterfactual view therefore serves as a consistency check, retaining the token-level visual credit signal only when the privileged visual difference agrees with the corresponding evidence-present versus evidence-suppressed change.

3.4 Rollout-level Visual Credit Calibration

For each image-question pair (x, q), the policy samples G responses $\{y^{(i)}\}_{i=1}^{G}$. Although the verified token-level visual credit captures token-wise evidence within each response, sibling rollouts may differ substantially in visual reliance. We therefore exploit this within-group variation to calibrate visual credit while preserving its token-level structure.

For each response $y^{(i)}$ of length T_i, we summarize its overall visual reliance by averaging the verified token-level visual credit:

$$\bar{e}_i^{\mathrm{vis}} = \mathrm{mean}_t\left(\tilde{e}_{i,t}^{\mathrm{vis}}\right). \tag{5}$$

Let $\mathcal{G}^+ = \{i : A_{\mathrm{ORM}}^{(i)} > 0\}$ and $\mathcal{G}^- = \{i : A_{\mathrm{ORM}}^{(i)} < 0\}$ denote the correct and incorrect subsets. We define their midpoint and the normalized relative visual score of each response as:

$$c = \frac{1}{2}\left[\mathrm{mean}_{i\in\mathcal{G}^+}\left(\bar{e}_i^{\mathrm{vis}}\right) + \mathrm{mean}_{i\in\mathcal{G}^-}\left(\bar{e}_i^{\mathrm{vis}}\right)\right], \qquad s_i = \frac{\bar{e}_i^{\mathrm{vis}} - c}{\sum_{j=1}^{G}\left|\bar{e}_j^{\mathrm{vis}} - c\right|}. \tag{6}$$

The normalized relative visual score is mapped to a bounded rollout-level weight,

$$w_i = \mathrm{clip}\left(1 + \beta \tanh(s_i), w_{\min}, w_{\max}\right), \qquad e_{i,t} = w_i \cdot \tilde{e}_{i,t}^{\mathrm{vis}}. \tag{7}$$

Notably, when either subset is empty, we set $w_i = 1$ for all responses in the group. The rollout-level weight only rescales the magnitude of the token-level visual credit; it does not alter its token-wise structure or determine the optimization direction.

3.5 Reward-Anchored Visual Credit Assignment

The calibrated visual credit $e_{i,t}$ refines token-level credit assignment, but cannot alone determine whether to reinforce or penalize a response. We therefore anchor the optimization direction to outcome advantage $A_{\mathrm{ORM}}^{(i)}$ and use visual credit only to modulate its magnitude at selected tokens.

Sign-Preserving Advantage Modulation. Before modulation, we center the visual credit within each response and map it to a bounded token-level adjustment:

$$\hat{e}_{i,t} = e_{i,t} - \mathrm{mean}_s\left(e_{i,s}\right), \qquad r_{i,t} = \lambda \tanh\left(\hat{e}_{i,t}\right), \qquad m_{i,t} = \mathbb{I}\left(|r_{i,t}| > \delta\right), \tag{8}$$

where λ controls the modulation strength and δ filters weak visual credit adjustments.

For tokens selected by the mask, we incorporate the visual adjustment into the outcome advantage while preserving its original sign:

$$\tilde{A}_{i,t} = \begin{cases} \max\left(A^{(i)}_{\mathrm{ORM}} + r_{i,t},\, 0\right), & A^{(i)}_{\mathrm{ORM}} \geq 0, \\ \min\left(A^{(i)}_{\mathrm{ORM}} + r_{i,t},\, 0\right), & A^{(i)}_{\mathrm{ORM}} < 0. \end{cases} \tag{9}$$

Thus, the outcome reward determines the optimization direction, while the visual credit refines token-level updates by strengthening or weakening their magnitude without reversing their sign.

Dual-Path Policy Optimization. Inspired by prior work Pan et al. (2026), we retain the original outcome advantage for unmodulated tokens and apply the visually modulated advantage only to selected tokens. Let

$$\mathcal{U}_i = \{t : m_{i,t} = 0\}, \qquad \mathcal{M}_i = \{t : m_{i,t} = 1\}, \tag{10}$$

denote the unmodulated and modulated token sets, respectively. For a token-level advantage A, we define the PPO clipped surrogate as:

$$\ell_{i,t}(A;\theta) = \min\left(\rho_{i,t}(\theta)A,\ \mathrm{clip}\left(\rho_{i,t}(\theta), 1-\epsilon, 1+\epsilon\right)A\right), \tag{11}$$

where

$$\rho_{i,t}(\theta) = \frac{\pi_\theta(y^{(i)}_t \mid x, q, y^{(i)}_{<t})}{\pi_{\theta_{\mathrm{old}}}(y^{(i)}_t \mid x, q, y^{(i)}_{<t})}. \tag{12}$$

The two optimization paths are normalized independently within each response:

$$\mathcal{J}^{(i)}(\theta) = \frac{1}{|\mathcal{U}_i|}\sum_{t\in\mathcal{U}_i} \ell_{i,t}\left(A^{(i)}_{\mathrm{ORM}};\theta\right) + \eta\frac{1}{|\mathcal{M}_i|}\sum_{t\in\mathcal{M}_i} \ell_{i,t}\left(\tilde{A}_{i,t};\theta\right). \tag{13}$$

Independent normalization prevents the contribution of sparse visually credited tokens from being diluted by the substantially larger unmodulated token set. The final objective averages $\mathcal{J}^{(i)}$ over the rollout group together with a KL regularizer to a frozen reference policy.

4 Experiments

4.1 Experimental Setup

Training Data and Benchmarks. We train VisCA on the same 6.2K synthetic visual question-answering samples released by Vision-OPD (Yuan et al., 2026). We evaluate fine-grained visual understanding on V* Bench (Wu & Xie, 2024), ZoomBench (Wei et al., 2026), HR-Bench (4K/8K) (Wang et al., 2025), and MME-RealWorld (EN/CN) (Zhang et al., 2025b). Generalization is further evaluated on the held-out MMVP (Tong et al., 2024b), CV-Bench (Tong et al., 2024a), and MMStar (Chen et al., 2024) benchmarks.

Models and Baselines. We use Qwen3.5 (4B and 9B) (Qwen Team, 2026) as base models for our VisCA and compare our models against two groups of baselines:

1. **State-of-the-Art MLLMs:** (a) "Thinking-with-Images" agentic models that perform multi-step reasoning while iteratively cropping and zooming into relevant image regions, including Thyme-RL (Zhang et al., 2025a), DeepEyesV2 (Hong et al., 2026), and SenseNova-MARS (Chng et al., 2025); (b) closed-source models, including Gemini-3-Flash (Google DeepMind, 2025), GPT-5.2 (OpenAI, 2025), and Doubao-Seed-2.0-Pro (ByteDance Seed Team, 2026); and (c) open-source models spanning a range of model scales, including Qwen3-VL-Instruct (Bai et al., 2025), ZwZ (Wei et al., 2026), and Qwen3.5 (Qwen Team, 2026).
2. **Alternative Training Strategies:** (a) GRPO (Shao et al., 2024), which applies group relative policy optimization with outcome rewards; (b) OPSD (Zhao et al., 2026), an on-policy self-distillation method that matches the student to an answer-hint teacher via forward KL; (c) RLSD (Yang et al., 2026), which uses the GRPO outcome advantage to determine the update direction and the teacher log-probability gap between answer-hint and original contexts to scale its magnitude; and (d) Vision-OPD (Yuan et al., 2026), which distills regional perception from a crop-conditioned teacher to a full-image student via Jensen–Shannon divergence.

Table 1: Comparison on six fine-grained visual benchmarks. We report accuracy (%) for each model. Among open-source models, the best are highlighted in **bold**, and the second-best are underlined.

Model	Param Size	V* Bench	ZoomBench	HR-Bench 4K	HR-Bench 8K	MME-RW EN	MME-RW CN	Average$_6$
"Thinking-with-Images" Agentic Models								
Thyme-RL	7B	83.77	45.80	77.38	70.50	64.94	64.64	67.84
DeepEyesV2	7B	85.34	44.50	74.25	73.12	61.92	63.92	67.18
SenseNova-MARS	8B	89.53	55.38	82.12	80.50	69.74	70.47	74.62
Closed-Source Models								
Gemini-3-Flash	-	85.86	61.30	88.25	85.00	73.35	72.05	77.64
GPT-5.2	-	85.34	54.56	84.75	78.50	71.83	67.48	73.74
Doubao-Seed-2.0-Pro	-	89.53	59.41	84.50	77.25	70.80	65.88	74.56
Open-Source Models								
Qwen3-VL-Instruct	4B	82.72	45.09	79.12	75.00	63.31	63.11	68.06
ZwZ	4B	89.01	55.86	83.00	78.38	66.90	67.64	73.47
Qwen3.5	4B	83.77	49.47	85.62	79.88	62.71	62.67	70.69
Qwen3-VL-Instruct	8B	83.77	43.31	78.25	74.12	63.24	64.39	67.85
ZwZ	8B	87.96	56.92	84.25	81.75	66.52	68.04	74.24
Qwen3.5	9B	84.82	52.07	84.25	80.12	70.25	67.33	73.14
Qwen3-VL-Instruct	235B	88.48	55.38	83.25	79.62	70.96	68.90	74.43
Qwen3.5	397B	84.29	57.28	86.50	84.25	73.68	67.58	75.60
VisCA (Ours)	4B	92.15	60.47	**87.25**	82.62	72.69	71.18	77.73
VisCA (Ours)	9B	**96.34**	**60.59**	**87.25**	**85.25**	**73.82**	**72.82**	**79.35**

Implementation Details. All methods share the same training data and optimization settings. We construct the evidence-suppressed counterfactual view x^- from the evidence-present privileged view x^+ by $0.1\times$ bilinear downsampling followed by nearest-neighbor upsampling, and use a batch size of 64, 8 rollouts per prompt, and a learning rate of 2×10^{-6}. *More implementation details are provided in Appendix A.*

4.2 Performance on Fine-Grained Visual Tasks

VisCA achieves strong fine-grained visual performance against diverse MLLM baselines. Table 1 compares VisCA with representative multimodal models, including 'Thinking-with-Images" agentic models, closed-source models, and open-source models across different scales. With Qwen3.5-4B and Qwen3.5-9B backbones, VisCA achieves average scores of 77.73% and 79.35% across six fine-grained visual benchmarks, respectively. At the 4B scale, VisCA improves over the original Qwen3.5-4B baseline by 7.04 points and achieves competitive performance with substantially larger open-source models, including Qwen3.5-397B (75.60%) and Qwen3-VL-235B-Instruct (74.43%). At the 9B scale, VisCA further improves over Qwen3.5-9B by 6.21 points and surpasses several strong closed-source models, including Gemini-3-Flash (77.64%) and GPT-5.2 (73.74%). These results demonstrate that VisCA effectively internalizes fine-grained visual understanding as an intrinsic capability of the model.

VisCA consistently improves over representative post-training baselines across model scales. Table 2 compares VisCA with representative post-training approaches using the same backbone and training data at each model scale. At the 4B scale, VisCA achieves an average score of 77.73%, outperforming GRPO, OPSD, RLSD, and Vision-OPD by 3.45, 5.28, 5.11, and 3.11 points, respectively. At the 9B scale, VisCA reaches 79.35%, outperforming GRPO and Vision-OPD by 2.31 and 1.16 points, respectively. These results demonstrate that VisCA provides a more effective post-training strategy for fine-grained visual understanding than representative outcome-based reinforcement learning and privileged-view self-distillation methods.

4.3 Generalization to Held-Out Multimodal Tasks

To examine whether post-training for fine-grained visual understanding compromises broader multimodal capabilities, we further evaluate the post-trained checkpoints on MMVP, CV-Bench, and MMStar. As shown in Table 2, VisCA achieves average scores of 80.69% and 82.98% at the

Table 2: Comparison with existing methods on fine-grained visual tasks and held-out tasks. † denotes the use of officially released model weights.

Method	Fine-Grained Visual Tasks						Average$_6$	Held-out Tasks			Average$_3$
	V* Bench	ZoomBench	HR-Bench 4K	HR-Bench 8K	MME-RW EN	MME-RW CN		MMVP	CV-Bench	MMStar	
Qwen3.5-4B											
Vanilla	83.77	49.47	85.62	79.88	62.71	62.67	70.69	76.00	86.90	**74.07**	78.99
GRPO	87.43	60.12	82.12	79.38	68.20	68.41	74.28	72.67	84.46	68.53	75.22
OPSD	82.20	54.91	79.62	78.62	70.67	68.65	72.45	80.33	86.98	69.73	79.01
RLSD	81.15	58.70	79.38	77.75	69.88	68.84	72.62	74.33	84.49	68.40	75.74
Vision-OPD†	89.01	59.17	81.00	78.38	70.93	69.22	74.62	79.33	87.10	70.73	79.05
VisCA (Ours)	**92.15**	**60.47**	**87.25**	**82.62**	**72.69**	**71.18**	**77.73**	**81.67**	**87.53**	72.87	**80.69**
Qwen3.5-9B											
Vanilla	84.82	52.07	84.25	80.12	70.25	67.33	73.14	81.67	88.37	**77.27**	82.44
GRPO	91.62	62.01	82.75	82.88	72.00	70.98	77.04	79.00	87.92	75.73	80.88
OPSD	87.96	57.63	82.75	80.75	71.80	68.95	74.97	81.00	88.13	71.93	80.35
RLSD	91.10	60.00	84.00	81.12	71.91	70.66	76.47	80.00	88.42	74.00	80.81
Vision-OPD†	91.10	**64.50**	86.50	84.88	71.82	70.34	78.19	82.67	88.17	75.80	82.21
VisCA (Ours)	**96.34**	60.59	**87.25**	**85.25**	**73.82**	**72.82**	**79.35**	**83.67**	**88.87**	76.40	**82.98**

4B and 9B scales, respectively, exceeding the corresponding pretrained models by 1.70 and 0.54 points. VisCA also consistently outperforms Vision-OPD across all three held-out benchmarks at both model scales, improving the average score by 1.64 points at 4B and 0.77 points at 9B. These results indicate that the gains in fine-grained visual understanding obtained by VisCA do not come at the expense of broader multimodal capabilities.

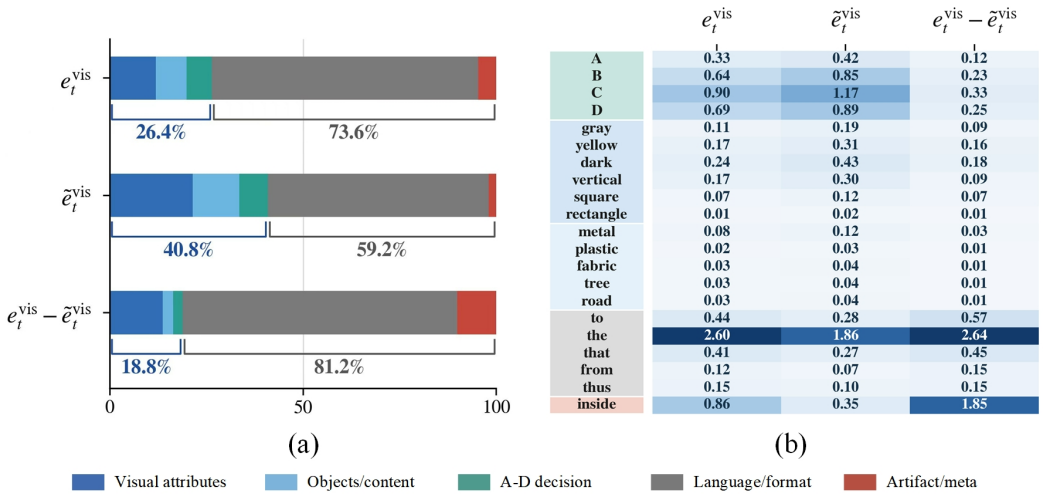

Figure 4: Semantic analysis of token-level visual credit. (a) Relative top-5 composition of e_t^{vis}, $\tilde{e}_t^{\text{vis}}$, and $e_t^{\text{vis}} - \tilde{e}_t^{\text{vis}}$ after excluding Other. (b) Representative token-level component scores.

4.4 Analysis of Visual Credit Assignment

Inspired by the semantic attribution analysis in VAD (Zhang et al., 2026), we examine whether counterfactual verification produces more reliable token-level visual credit by comparing the privileged visual difference e_t^{vis}, the verified signal $\tilde{e}_t^{\text{vis}}$, and the filtered component $e_t^{\text{vis}} - \tilde{e}_t^{\text{vis}}$ in Figure 4.

Counterfactual verification concentrates visual credit on visually grounded and decision-relevant tokens. As shown in Figure 4(a), visual attributes, objects/content, and A–D decision tokens jointly account for 40.8% of the verified visual credit $\tilde{e}_t^{\text{vis}}$, compared with 26.4% in the original privileged difference e_t^{vis}. In contrast, these categories constitute only 18.8% of the filtered component $e_t^{\text{vis}} - \tilde{e}_t^{\text{vis}}$, which is dominated by language/format and artifact/meta tokens (81.2%). This redistribution suggests that the counterfactual consistency check preferentially retains token differences associated with task-relevant visual evidence while filtering differences dominated by linguistic or transformation-related effects.

Token-level scores exhibit the same semantic separation. Figure 4(b) shows that the verified signal assigns stronger credit to answer choices and representative visual concepts. For example,

Table 3: Ablation of key design choices in VisCA.

Variant	V* Bench	ZoomBench	HR-Bench 4K	HR-Bench 8K	MME-RW EN	MME-RW CN	Average$_6$
Base	85.34	**63.43**	78.50	74.25	69.78	63.95	72.54
+Visual Evidence	90.58	59.41	85.50	81.75	71.35	69.02	76.27
+Visual Evidence −Dual-Path	85.34	59.76	79.75	78.88	68.50	67.30	73.26
+Visual Evidence +Verification	<u>92.15</u>	59.88	85.62	80.75	71.99	**71.51**	<u>76.98</u>
+Visual Evidence +Calibration	**93.19**	56.80	85.75	81.75	<u>72.31</u>	70.42	76.70
VisCA (Full)	<u>92.15</u>	<u>60.47</u>	**87.25**	<u>82.62</u>	**72.69**	71.18	**77.73**
$-A_{\mathrm{ORM}}$ Anchoring	88.48	58.46	<u>86.50</u>	**83.12**	71.84	<u>71.40</u>	76.63
−Token-Level Granularity	88.48	58.58	83.62	80.38	71.40	70.97	75.57

the scores for A–D increase from 0.33/0.64/0.90/0.69 under e_t^{vis} to 0.42/0.85/1.17/0.89 under $\tilde{e}_t^{\mathrm{vis}}$, while visual tokens such as "*yellow*", "*dark*", and "*vertical*" exhibit similar increases. By contrast, language-oriented tokens such as "*the*", "*to*", and "*inside*" are substantially more prominent in the filtered component. Together, these results support the intended role of counterfactual verification. Privileged visual differences provide candidate evidence, while the matched evidence-suppressed view helps retain signals more closely associated with fine-grained visual evidence.

4.5 Ablation of Key Design Choices

As shown in Table 3, we ablate the key design choices of VisCA using Qwen3.5-4B, with all variants using the same training data and recipe. *More ablations and analyses are provided in the Appendix.*

Sparse token-level visual evidence benefits from dual-path optimization. The *Base* model achieves an average score of 72.54. Introducing the raw privileged visual difference e_t^{vis} as token-level visual evidence (+*Visual Evidence*) improves the average to 76.27. Removing dual-path optimization while retaining the same visual evidence (+*Visual Evidence* −*Dual-Path*) reduces the average to 73.26. This result is consistent with the sparse token-level visual dependency observed in Figure 2: independently normalizing the modulated and unmodulated paths prevents sparse visually modulated tokens from being overwhelmed by the much larger unmodulated token set, thereby preserving their contribution to optimization.

Counterfactual verification and rollout-level calibration provide complementary gains. Building on raw token-level visual evidence, counterfactual verification (+*Visual Evidence* +*Verification*) improves the average score from 76.27 to 76.98, while rollout-level calibration (+*Visual Evidence* +*Calibration*) improves it to 76.70. Combining both components yields the full VisCA model with an average score of 77.73, outperforming either component alone. These gains indicate that counterfactual verification improves visual credit by selectively retaining visually grounded and decision-relevant token differences, while rollout-level calibration further refines it by adjusting credit magnitude across sibling rollouts.

Visual credit should refine reward-derived credit rather than replace it. Removing A_{ORM} anchoring ($-A_{\mathrm{ORM}}$ *Anchoring*) reduces the average score from 77.73 to 76.63, showing that visual credit is more effective when it modulates update magnitude while the verifier-derived advantage preserves the optimization direction. Replacing token-level visual credit $e_{i,t}$ with a single rollout-level relative visual score s_i shared by all tokens in the response (−*Token-Level Granularity*) further decreases the average score to 75.57. These results show that rollout-level information should calibrate token-level visual credit rather than replace its fine-grained token-wise structure.

5 Conclusion

We introduced VisCA, a self-distilled reinforcement learning framework for fine-grained visual credit assignment. The core idea is to treat privileged visual differences as evidence for refining reward-derived credit rather than as credit signals themselves. Through counterfactual verification and rollout-level calibration, VisCA converts privileged visual differences into verified fine-grained credit while preserving the reward-determined optimization direction. Experiments demonstrate consistent improvements at both 4B and 9B scales over representative post-training methods, while maintaining strong performance on held-out multimodal tasks.

LLM USAGE STATEMENT

During the preparation of this work, the authors used large language models (LLMs), including ChatGPT, as assistive tools for proofreading, grammatical correction, language refinement, and improving the readability and clarity of the manuscript. All technical content and the final manuscript were carefully reviewed and verified by the authors.

REPRODUCIBILITY STATEMENT

We provide detailed descriptions of the VisCA framework and experimental setup, including key training configurations and hyperparameters, throughout the paper. All datasets and benchmarks used for training and evaluation are publicly available. Our implementation builds upon the publicly available Vision-OPD codebase, with VisCA developed within the same training framework. The code and model checkpoints developed for this work are not included in the current submission, as their public release is subject to our company's internal review and approval process. We intend to release these artifacts upon completion of the required internal approval process to facilitate full reproduction of our results.

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 *International Conference on Learning Representations*, volume 2024, pp. 21246–21263, 2024.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025.

Jinhe Bi, Peng Liao, Zengjie Jin, Volker Tresp, Fei Shen, Yunpu Ma, Tat-Seng Chua, et al. Opd-v: Visual on-policy self-distillation with modality balance. *arXiv preprint arXiv:2608.05131*, 2026.

ByteDance Seed Team. Seed2.0 model card: Towards intelligence frontier for real-world complexity, February 2026. URL https://lf3-static.bytednsdoc.com/obj/eden-cn/lapzild-tss/ljhwZthlaukjlkulzlp/seed2/0214/Seed2.0%20Model%20Card.pdf.

Yishuo Cai, Jiahui Liu, Yuanxin Liu, Haobo Deng, Linli Yao, Yuhao Zheng, Kun Ouyang, Zhimo Li, Ziyue Wang, Xu Sun, et al. Thinking without images: Internalizing visual manipulation with on-policy self-distillation. *arXiv preprint arXiv:2606.08719*, 2026.

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? *Advances in Neural Information Processing Systems*, 37:27056–27087, 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.

Ganqu Cui, Lifan Yuan, Zefan Wang, Hanbin Wang, Yuchen Zhang, Jiacheng Chen, Wendi Li, Bingxiang He, Yuchen Fan, Tianyu Yu, et al. Process reinforcement through implicit rewards. *arXiv preprint arXiv:2502.01456*, 2025.

Google DeepMind. Gemini 3 Flash: Model card, December 2025. URL https://deepmind.google/models/model-cards/gemini-3-flash/.

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.

Ahmed Heakl, Abdelrahman M Shaker, Youssef Mohamed, Rania Elbadry, Omar Fetouh, Fahad Shahbaz Khan, and Salman Khan. Cepo: Rlvr self-distillation using contrastive evidence policy optimization. *arXiv preprint arXiv:2605.19436*, 2026.

Jack Hong, Chenxiao Zhao, ChengLin Zhu, Weiheng Lu, and Guohai Xu. Deepeyesv2: Toward agentic multimodal model. In *International Conference on Learning Representations*, volume 2026, pp. 114851–114872, 2026.

Wenyi Hong, Wenmeng Yu, Xiaotao Gu, Guo Wang, Guobing Gan, Haomiao Tang, Jiale Cheng, Ji Qi, Junhui Ji, Lihang Pan, et al. Glm-4.5 v and glm-4.1 v-thinking: Towards versatile multimodal reasoning with scalable reinforcement learning. *arXiv preprint arXiv:2507.01006*, 2025.

Jonas Hübotter, Frederike Lübeck, Lejs Behric, Anton Baumann, Marco Bagatella, Daniel Marta, Ido Hakimi, Idan Shenfeld, Thomas Kleine Buening, Carlos Guestrin, et al. Reinforcement learning via self-distillation. *arXiv preprint arXiv:2601.20802*, 2026.

Mahyar Khayatkhoei, Prateek Chhikara, Filip Ilievski, et al. Mllms know where to look: Training-free perception of small visual details with multimodal llms. In *International Conference on Learning Representations*, volume 2025, pp. 68194–68213, 2025.

Yang Li, Zhichen Dong, Yuhan Sun, Weixun Wang, Shaopan Xiong, Yijia Luo, Jiashun Liu, Han Lu, Jiamang Wang, Wenbo Su, et al. Attention illuminates llm reasoning: The preplan-and-anchor rhythm enables fine-grained policy optimization. *arXiv preprint arXiv:2510.13554*, 2025.

Yijiang Li, Yijun Liang, Yunjie Tian, Bingyang Wang, Ke Zhang, Zhenfei Yin, Di Fu, Philip Torr, and Nuno Vasconcelos. Self-supervised visual on-policy distillation. *arXiv preprint arXiv:2608.14144*, 2026a.

Ziheng Li, Liu Kang, Feng Xiao, Luxi Xing, Qingyi Si, Zhuoran Li, Weikang Gong, Deqing Yang, Yanghua Xiao, and Hongcheng Guo. Outcome-grounded advantage reshaping for fine-grained credit assignment in mathematical reasoning. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics*, pp. 24681–24693, 2026b.

Hunter Lightman, Vineet Kosaraju, Yuri Burda, Harrison Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's verify step by step. In *International Conference on Learning Representations*, volume 2024, pp. 39578–39601, 2024.

Peng Liu, Haozhan Shen, Chunxin Fang, Zhicheng Sun, Jiajia Liao, and Tiancheng Zhao. Vlm-fo1: Bridging the gap between high-level reasoning and fine-grained perception in vlms. *arXiv preprint arXiv:2509.25916*, 2025.

Liangchen Luo, Yinxiao Liu, Rosanne Liu, Samrat Phatale, Meiqi Guo, Harsh Lara, Yunxuan Li, Lei Shu, Yun Zhu, Lei Meng, et al. Improve mathematical reasoning in language models by automated process supervision. *arXiv preprint arXiv:2406.06592*, 2024.

OpenAI. Introducing gpt-5.2, December 2025. URL `https://openai.com/index/introducing-gpt-5-2/`.

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.

Qwen Team. Qwen3.5: Towards native multimodal agents, February 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, YK Li, Yang Wu, et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

Shengbang Tong, Ellis Brown, Penghao Wu, Sanghyun Woo, Manoj Middepogu, Sai C Akula, Jihan Yang, Shusheng Yang, Adithya Iyer, Xichen Pan, et al. Cambrian-1: A fully open, vision-centric exploration of multimodal llms. *Advances in Neural Information Processing Systems*, 37:87310–87356, 2024a.

Shengbang Tong, Zhuang Liu, Yuexiang Zhai, Yi Ma, Yann LeCun, and Saining Xie. Eyes wide shut? exploring the visual shortcomings of multimodal llms. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9568–9578, 2024b.

Haochen Wang, Yuhao Wang, Tao Zhang, Yikang Zhou, Yanwei Li, Jiacong Wang, Ye Tian, Jiahao Meng, Zilong Huang, Guangcan Mai, et al. Grasp any region: Towards precise, contextual pixel understanding for multimodal llms. In *International Conference on Learning Representations*, volume 2026, pp. 2222–2250, 2026a.

Peiyi Wang, Lei Li, Zhihong Shao, Runxin Xu, Damai Dai, Yifei Li, Deli Chen, Yu Wu, and Zhifang Sui. Math-shepherd: Verify and reinforce llms step-by-step without human annotations. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*, pp. 9426–9439, 2024.

Shenzhi Wang, Le Yu, Chang Gao, Chujie Zheng, Shixuan Liu, Rui Lu, Kai Dang, Xiong-Hui Chen, Jianxin Yang, Zhenru Zhang, et al. Beyond the 80/20 rule: High-entropy minority tokens drive effective reinforcement learning for llm reasoning. *Advances in Neural Information Processing Systems*, 38:115452–115486, 2026b.

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, et al. 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. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13084–13094. IEEE, 2024.

Can Xie, Ruotong Pan, Xiangyu Wu, Zhang Yunfei, Jiayi Fu, Tingting Gao, and Guorui Zhou. Unlocking exploration in rlvr: Uncertainty-aware advantage shaping for deeper reasoning. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 19057–19076, 2026.

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.

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.

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, Xingyu Lu, Shukang Yin, Chaoyou Fu, Wei Chen, Xiao Hu, Bin Wen, Kaiyu Jiang, Changyi Liu, Tianke Zhang, et al. Thyme: Think beyond images. *arXiv preprint arXiv:2508.11630*, 2025a.

YiFan Zhang, Huanyu Zhang, Haochen Tian, Chaoyou Fu, Shuangqing Zhang, Junfei Wu, Feng Li, Kun Wang, Qingsong Wen, Zhang Zhang, et al. Mme-realworld: Could your multimodal llm challenge high-resolution real-world scenarios that are difficult for humans? In *International Conference on Learning Representations*, volume 2025, pp. 89655–89701, 2025b.

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.

Appendix

The Supplementary Material of the paper is organized as:

- **Appendix A:** Implementation Details.
- **Appendix B:** Ablation Studies on Hyper-parameters.
- **Appendix C:** Ablation Studies on Counterfactual Views.
- **Appendix D:** Token-Level Qualitative Analysis of Credit Assignment.
- **Appendix E:** Qualitative Case Studies.
- **Appendix F:** Computational Cost Analysis.
- **Appendix G:** Limitations and Future Work.

A Implementation Details

A.1 Benchmark Details

We evaluate VisCA on two groups of benchmarks covering fine-grained visual understanding and broader multimodal capabilities.

Fine-grained visual understanding. V* Bench (Wu & Xie, 2024) evaluates the ability to localize and recognize small visual targets embedded in complex scenes. ZoomBench (Wei et al., 2026) focuses on visual question answering that requires resolving fine-grained details at different zoom levels. HR-Bench (Wang et al., 2025) evaluates high-resolution visual perception at both 4K and 8K resolutions. MME-RealWorld (Zhang et al., 2025b) assesses perception and reasoning over high-resolution images from diverse real-world scenarios, with both English (EN) and Chinese (CN) subsets. Together, these benchmarks cover complementary challenges involving small objects, subtle visual details, high-resolution perception, and real-world visual reasoning.

Held-out multimodal tasks. We further evaluate the post-trained models on MMVP (Tong et al., 2024b), CV-Bench (Tong et al., 2024a), and MMStar (Chen et al., 2024). These benchmarks cover complementary aspects of general multimodal understanding and are not used for VisCA training. Their performance is used to assess whether improving fine-grained visual understanding comes at the expense of broader multimodal capabilities.

A.2 Implementation Details

All experiments are conducted on 8 H20 GPUs with full-parameter training. Each example provides a full-image view x with the evidence region marked by a red bounding box, an evidence-present privileged view x^+, and an evidence-suppressed counterfactual view x^-. The evidence-suppressed view is obtained by applying $0.1\times$ bilinear downsampling followed by nearest-neighbor upsampling to the evidence-present crop. We use a training batch size of 64, 8 rollouts per prompt, and a learning rate of 2×10^{-6}. We set the modulation strength to $\lambda = 0.2$, the dual-path weighting coefficient to $\eta = 1.0$, the modulation threshold to $\delta = 0.05$ for Qwen3.5-4B and $\delta = 0.10$ for Qwen3.5-9B, and the rollout-level reweighting coefficient to $\beta = 0.5$. All methods share the same training data and optimization settings, with a maximum generation length of 1,024 and a single training epoch.

B Ablation Studies on Hyper-parameters

Rollout-level reweighting coefficient β. The coefficient β controls the strength of rollout-level calibration by determining how strongly the relative visual reliance of sibling rollouts rescales their verified token-level visual credit. As shown in Figure 5(a), increasing β from 0.10 to 0.50 steadily improves the average performance from 76.73 to 77.73, suggesting that appropriately differentiating

sibling rollouts according to their relative visual reliance provides more informative credit calibration than near-uniform weighting. Further increasing β to 0.75 and 1.00 reduces the average score to 77.47 and 76.77, respectively, indicating that overly strong rollout-level reweighting can over-amplify differences in visual reliance across sampled responses. We therefore set $\beta = 0.50$, which provides a balanced calibration of visual credit while preserving its token-level structure.

Modulation strength λ. The coefficient λ controls the magnitude of the token-level adjustment introduced by the calibrated visual credit, and thus determines how strongly visual evidence refines the verifier-derived outcome advantage. As shown in Figure 5(b), performance improves from 76.17 at $\lambda = 0.10$ to 77.73 at $\lambda = 0.20$, but decreases to 76.78 and 76.14 when λ is further increased to 0.25 and 0.30, respectively. This trend suggests that moderate modulation is sufficient to emphasize visually informative token-level decisions, whereas overly strong modulation can place excessive weight on the auxiliary visual credit. We therefore use $\lambda = 0.20$, allowing visual credit to meaningfully refine update magnitudes while keeping the optimization anchored to the verifier-derived outcome advantage.

Modulation threshold δ. The threshold δ determines which token-level visual adjustments are sufficiently strong to enter the visually modulated optimization path, filtering out weak adjustments with limited modulation strength. As shown in Figure 5(c), using no threshold yields an average score of 76.18, while introducing a small threshold improves performance to 76.34 at $\delta = 0.02$ and further to 77.73 at $\delta = 0.05$. This improvement suggests that excluding weak visual-credit adjustments helps concentrate modulation on more informative token-level signals. Increasing the threshold beyond this point leads to lower performance, with scores of 77.33 and 76.28 at $\delta = 0.10$ and 0.15, respectively, suggesting that an overly restrictive threshold may exclude useful visually grounded adjustments from the modulated path. We therefore set $\delta = 0.05$, balancing the exclusion of weak adjustments against the retention of informative visual credit.

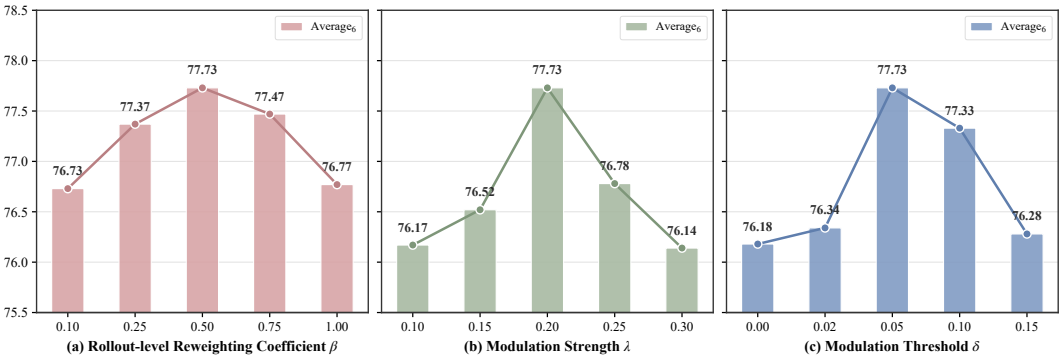

Figure 5: Ablation studies of (a) rollout-level reweighting coefficient β, (b) modulation strength λ, and (c) modulation threshold δ. Results report the average performance across six benchmarks. All hyperparameter ablations in this section are conducted with Qwen3.5-4B.

C Ablation Studies on Counterfactual Views

Table 4 examines how the construction of the evidence-suppressed view x^- affects counterfactual verification. The results show that an effective counterfactual is not obtained simply by introducing a stronger visual intervention alone. Removing the image entirely or using a random crop yields Average_6 scores of 76.80 and 76.17, respectively. Although these constructions substantially alter the available visual information, they also change the visual condition beyond the task-relevant fine-grained evidence, making the resulting contrast less controlled. Masking performs better at 77.02 by applying a more localized intervention, but still underperforms the matched degradation used by VisCA. This suggests that evidence suppression is preferable to directly removing or replacing the relevant visual content, as the former better preserves the regional structure shared with the evidence-present view x^+.

The degradation level further reveals a trade-off between suppressing decisive details and preserving the matched visual condition. The $0.05\times$ setting achieves an Average_6 of 77.42, while the interme-

Table 4: Ablation study on different constructions of the counterfactual view x^-. We compare no image, random crop, masking, and downsampling-induced blur at different ratios, and report performance across six fine-grained visual benchmarks.

Counterfactual View	V* Bench	ZoomBench	HR-Bench 4K	HR-Bench 8K	MME-RW EN	MME-RW CN	Average$_6$
No Image	88.48	60.00	86.00	82.62	72.51	71.18	76.80
Random Crop	88.48	58.82	85.12	81.12	72.55	70.91	76.17
Mask	89.01	57.28	**87.25**	83.75	72.60	**72.25**	77.02
Blur$_{0.05}$	91.10	59.17	85.88	**83.88**	**72.93**	71.57	77.42
Blur$_{0.10}$	**92.15**	60.47	**87.25**	82.62	72.69	71.18	**77.73**
Blur$_{0.25}$	88.48	**60.83**	84.62	83.25	71.38	70.81	76.56

diate $0.10\times$ setting reaches the best performance of 77.73. The performance decreases to 76.56 at $0.25\times$. These results suggest that excessive degradation may remove information beyond the fine-grained evidence of interest, whereas insufficient degradation may leave too much answer-relevant detail available in x^-. The $0.10\times$ setting provides a better balance between these two effects, suppressing fine-grained evidence while preserving the region and its broader visual structure. Taken together, these results highlight the importance of constructing a matched counterfactual in VisCA. The goal is not to maximize the difference between x^+ and x^-, but to provide a controlled contrast in visual evidence, so that agreement between the privileged and counterfactual differences more reliably reflects sensitivity to the relevant fine-grained visual evidence.

D Token-Level Qualitative Analysis of Credit Assignment

Figure 6 visualizes the token-level credit assignment of different methods on representative rollouts. Compared with the self-distilled RLVR baseline RLSD (Yang et al., 2026), VisCA produces a more visually grounded credit distribution. High-credit regions are concentrated on tokens that encode the decisive fine-grained visual evidence and the corresponding answer decisions, whereas language-dominant tokens and generic reasoning expressions receive substantially weaker credit.

For example, VisCA assigns stronger credit to visual attributes, object-specific information, and answer tokens that directly depend on the image evidence, while avoiding excessive credit on surrounding linguistic scaffolding. In contrast, the baseline credit is more diffusely distributed across the response and does not distinguish visually consequential tokens as clearly. These qualitative examples are consistent with the semantic analysis in Section 4.4, showing that VisCA provides more precise token-level credit assignment by better localizing the learning signal to decisions grounded in fine-grained visual evidence.



How many small dark animals are visible on the grass?
GT: D.2

Standard RLSD

Looking at the area inside the red bounding box — which highlights the grass y strip between the road and the monument base — we can see a few small dark shapes · Upon closer inspection , these shapes appear to be two small birds or animals (possibly pige ons or similar ground -d w elling birds) sitting on the grass near the left side of the highlighted area · Thus , the correct count is 2 D · 2

VisCA (Ours)

Looking at the area inside the red bounding box — which highlights the grass y strip between the road and the monument base — we can see a few small dark shapes · Upon closer inspection , these shapes appear to be two small birds or animals (possibly pige ons or similar ground -d w elling birds) sitting on the grass near the left side of the highlighted area · Thus , the correct count is 2 D · 2

Figure 6: Token-level credit assignment of "VisCA" and the self-distilled RLVR baseline "RLSD" on a representative rollout. Green indicates higher credit, while red indicates lower credit.

Case1: Small Text / Number Reading:		Case2: Numeral Type / Anti-language Prior:	
			
User	What is the number written on the side of the blue vehicle? (A) C519 (B) C518 (C) C517 (D) C520	User	What type of numerals are used on the clock face? (A) Roman numerals (B) Arabic numerals (C) Chinese numerals (D) Binary numerals
Qwen3.5	Looking at the image, we can see a blue bus on the left side of the frame. On its side, near the rear wheel and above the checkered pattern, there is a white number written in black. That number is clearly C519. Correct Answer: A. C519	Qwen3.5	The clock face on the Elizabeth Tower (commonly known as Big Ben) in the image displays time using Arabic numerals—specifically, the numbers 1 through 12 are shown in a circular arrangement around the clock face, which is characteristic of modern Western clocks. Although some traditional clocks use Roman numerals, this one clearly uses digits like 1, 2, ..., 12—which are Arabic numerals. Correct Answer: B. Arabic numerals
VisCA	The number clearly written next to the checkered pattern on the rear wheel well of the blue bus is: C 518 Correct option: B. C518 Answer: B	VisCA	The clock face uses Roman numerals (I, II, III, ..., XII) to indicate the hours. Correct option: A. Roman numerals Answer: A

Figure 7: Representative successful cases of VisCA on fine-grained visual understanding.

Failure Case1: Scene Layout:		Failure Case2: Pairwise Relation:	
			
User	Where are the houses located relative to the tea fields? (A) At the top right corner (B) In the bottom center (C) At the top left corner (D) In the middle of the tea fields	User	What is the relative position of the red car compared to the black car? (A) To the right of the black car (B) To the left of the black car (C) Behind the black car (D) In front of the black car
Qwen3.5	Looking at the image, The tea fields dominate the foreground and middle ground… In the background, especially toward the right side, we can see a cluster of small buildings or houses nestled among trees… These houses are not in the bottom center … nor at the top left or top right corners (those areas show hills, trees, or sky). They are situated within the broader landscape of tea fields, but more accurately, they are in the middle distance on the right … Best fit: (D) In the middle of the tea fields	Qwen3.5	Looking at the image, we can see a red car in the background on the far left side of the scene—it's partially visible behind trees and near a brick wall. There is no black car visible in the image at all. Since the black car is not present, any spatial relationship (left, right, front, behind) cannot be established. If we assume the "black car" is a hypothetical object to the right (which is common in such questions when one object is missing), then the red car would be to the left of it. Thus, the best answer among the choices, assuming the black car is implied to be elsewhere (e.g., to the right), is: (B) To the left of the black car . Final Answer: (B) To the left of the black car
VisCA	The cluster of houses with red-tiled roofs and a white tent is nestled among the tea fields on the right side of the image, roughly in the middle ground—not at the very top right corner (which has forested hills and misty mountains), nor at the top left … nor in the bottom center … (A) At the top right corner—partially correct, but misleading; they're more mid-right, not the extreme top corner. (D) In the middle of the tea fields—accurate … the houses are surrounded by and integrated into the sprawling tea plantations. Best answer: (D) In the middle of the tea fields	VisCA	The red car is clearly positioned to the left of the dark blue/black SUV parked just behind it near the brick wall and trees. Correct option: (B) To the left of the black car Answer: B

Figure 8: Representative failure cases of VisCA on fine-grained visual understanding.

E Qualitative Case Studies

Successful cases. Figure 7 presents two representative cases illustrating how VisCA improves the use of fine-grained visual evidence under standard full-image inference. In the first case, the base model misreads the small number on the blue vehicle as "C519," whereas VisCA correctly identifies it as "C518," reflecting improved sensitivity to small and localized visual details. The second case presents a different challenge in which the visual evidence conflicts with a plausible language prior. The base model follows this prior and incorrectly predicts that the clock face uses Arabic numerals, even generating a plausible explanation around the incorrect answer. In contrast, VisCA correctly recognizes the Roman numerals shown on the clock face, suggesting that it relies more effectively on the actual fine-grained visual evidence rather than the language prior.

Importantly, both cases are evaluated using only the original full image. The privileged and counterfactual views used by VisCA are available only during training and are not required at inference time. The improvements therefore do not depend on additional visual access at test time, suggesting that the fine-grained visual capability encouraged by privileged visual credit assignment is internalized into the model itself. Together with the quantitative improvements across fine-grained visual benchmarks, these examples provide qualitative evidence that VisCA helps the model better exploit small but decisive visual cues embedded in complex visual contexts.

Failure cases. Figure 8 presents representative failure cases that reveal a limitation of the current VisCA formulation. These examples involve spatial relationships whose decisive evidence cannot be localized to an isolated visual detail. The first case requires understanding the location of a group of houses within the broader scene layout, while the second requires establishing the relative position between two distinct objects. In both settings, resolving the answer depends not only on recognizing individual visual elements, but also on relating them within a broader spatial context.

F Computational Cost Analysis

Table 5 compares the computational cost of VisCA with representative RLVR and self-distillation methods. VisCA spends 22.22 s/step on teacher log-probability computation, substantially less than OPSD (46.24 s/step) and RLSD (86.20 s/step). This efficiency mainly stems from the form of privileged information and how the teacher signal is used. OPSD and RLSD condition the teacher on privileged textual solutions, such as reference reasoning traces, which lengthen the teacher context. In contrast, VisCA introduces privileged information directly through visual conditions, avoiding long textual privileged contexts. Moreover, unlike OPSD, which performs dense distribution matching over the vocabulary, VisCA only evaluates the probabilities of sampled tokens for visual credit assignment. As a result, although VisCA requires additional teacher forwards for counterfactual verification, its teacher-side computation remains relatively lightweight.

In terms of end-to-end training cost, VisCA takes 478.70 s per step, corresponding to a 13.7% overhead over GRPO (420.85 s/step), while remaining faster than OPSD (573.18 s/step) and RLSD (551.89 s/step). Vision-OPD incurs the lowest teacher-side cost (10.31 s/step), whereas VisCA requires additional visual-condition evaluations to verify whether privileged visual differences are attributable to fine-grained visual evidence. Overall, VisCA achieves more reliable visual credit assignment with a moderate computational overhead over standard RLVR, while maintaining favorable training efficiency compared with general self-distillation baselines.

Table 5: Average wall-clock time per training step (seconds) on fine-grained visual understanding task, using Qwen3.5-4B as the base model.

Method	Teacher log-prob (s/step)	Total step (s/step)
GRPO	–	420.85 ± 8.28
OPSD	46.24 ± 1.63	573.18 ± 11.26
RLSD	86.20 ± 2.27	551.89 ± 10.29
Vision-OPD	10.31 ± 1.52	372.29 ± 7.67
VisCA (Ours)	22.22 ± 2.42	478.70 ± 12.36

G LIMITATIONS AND FUTURE WORK

While our results demonstrate consistent improvements in fine-grained visual understanding, the current formulation of VisCA has several limitations that motivate further investigation.

Localized visual evidence. VisCA is primarily designed to assign token-level credit to fine-grained visual evidence localized within answer-relevant regions. This formulation is effective when the decisive information is concentrated in small text, subtle attributes, or localized object details, but is less suited to questions whose evidence is distributed across multiple regions. As illustrated by the failure cases in Figure 8, spatial-layout and multi-object relational questions require the model to jointly reason over several visual elements and their relationships within a broader spatial context. Improving the perception of an individual region may therefore be insufficient for learning these relational structures. Extending visual credit assignment from localized evidence to multi-region and relation-aware evidence is an important direction for future work.

Counterfactual view construction. The reliability of VisCA's visual credit depends on the construction of the evidence-present and evidence-suppressed views used during training. Our current implementation constructs a matched counterfactual by degrading fine-grained information within the same region, which provides a controlled contrast while preserving the broader visual structure. However, a single counterfactual view captures only one intervention on the underlying visual evidence and may not fully characterize more complex or compositional evidence. Future work could explore multiple complementary counterfactual views or adaptive evidence interventions to provide a richer basis for verifying token-level visual credit.