VC-RL: VISUAL CONTRAST-GUIDED REINFORCEMENT LEARNING FOR NATIVE VISUAL THINKING

Anonymous authors
Paper under double-blind review

ABSTRACT

Fine-grained visual understanding requires multimodal large language models to identify local details accurately and use them in visual reasoning. Native thinking can improve this capability, but the use of local visual feedback to improve long-form reasoning over full images remains underexplored. Local views improve a teacher's access to fine details while also changing image scale and context, so differences between teacher and student predictions can reflect both target evidence and other changes in visual conditions. We propose Visual Contrast-Guided Reinforcement Learning (VC-RL), which combines Group Relative Policy Optimization (GRPO) with visual on-policy self-distillation. The student generates thinking trajectories from full images, and a self-teacher scores the same trajectories under size-matched positive-evidence and counterfactual views constructed with shared preprocessing rules. The contrast between these predictions guides token selection and bounded modulation of group-relative advantages derived from sequence-level rewards, without reversing the signs of nonzero advantages. Regional visual feedback thereby contributes selectively to token-level optimization, while task-level rewards constrain the policy updates. Across six fine-grained and high-resolution visual evaluations, VC-RL achieves average accuracies of 79.21% and 81.33% with Qwen3.5-4B and Qwen3.5-9B, respectively, improving over the corresponding base models by 4.45 and 3.90 percentage points, with additional gains on multiple general multimodal benchmarks.

1 INTRODUCTION

Fine-grained visual understanding is fundamental to the ability of multimodal large language models (MLLMs) to perceive image details and perform visual reasoning. In high-resolution images, decisive clues often occupy only a small region, requiring models to identify local details and reason about them in relation to the question (Wei et al., 2026). Native thinking can improve performance on these tasks: as shown in Figure 1(a), enabling thinking yields higher accuracy for both the 4B and 9B base models across the evaluated visual benchmarks under otherwise matched evaluation settings. This observation highlights native thinking as an important avenue for improving fine-grained visual understanding and raises a further question: *How can post-training effectively improve native visual thinking?*

First, changes in teacher predictions induced by local views do not exclusively reflect useful visual feedback. Visual on-policy distillation uses local views to improve a teacher's access to fine-grained evidence and provides token-level supervision by scoring student-generated trajectories (Yuan et al., 2026; Aniri et al., 2026). At each student-generated prefix, reverse KL divergence (RKL) measures the discrepancy between student and teacher predictions and provides a token-level distillation objective. This discrepancy can reflect improved recognition of target-region details, but also changes in image scale and surrounding context. Figure 1(b) shows that thinking positions with high positive-teacher RKL can also exhibit high counterfactual-teacher RKL. Figure 1(c) further illustrates that generic reasoning and visual interpretation spans can both have high RKL despite differing in visual contrast strength. These observations motivate matched positive-evidence and counterfactual views: contrasting teacher predictions under the same student prefixes provides a regional contrast signal for selective feedback.

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

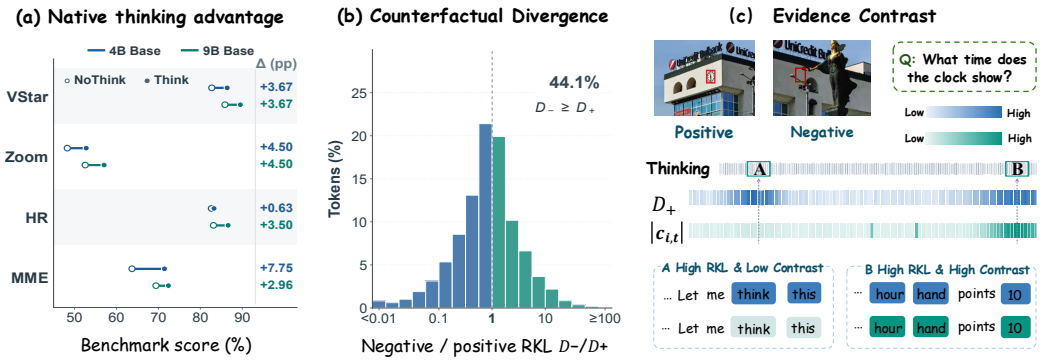


Figure 1: **Motivation for visual contrast in native visual thinking.** (a) Performance with and without native thinking. (b) Distribution of D_-/D_+ among the top 10% of thinking tokens ranked by D_+ within each question. D_+ and D_- denote student-to-teacher reverse KL divergences under positive-evidence and matched counterfactual views, evaluated at the same student prefixes. (c) Example spans with high D_+ but different regional contrast strengths. Here, $c_{i,t}$ is the positive-minus-negative teacher log-probability difference for the generated token; colors show D_+ and $|c_{i,t}|$.

Second, regional visual feedback alone is insufficient to determine the direction of policy updates. The self-teacher and student originate from the same base model, and the teacher's perceptual advantage comes primarily from privileged visual information rather than stronger reasoning capabilities. During native long-form thinking, the self-teacher provides feedback conditioned on student prefixes containing multiple intermediate judgments. Earlier errors may influence its predictions, leading it to favor continuations of incorrect reasoning even when clearer local evidence is available. Improved access to visual evidence therefore does not guarantee that teacher feedback aligns with the task objective, motivating the use of sequence-level task rewards to constrain policy optimization on fine-grained visual tasks.

To improve the use of fine-grained visual evidence during long-form thinking, we propose Visual Contrast-Guided Reinforcement Learning (VC-RL). The method combines Group Relative Policy Optimization (GRPO) (Shao et al., 2024) with visual on-policy self-distillation, using feedback from local regions to optimize thinking trajectories generated from full images. Because local views change both evidence visibility and other visual conditions, we construct size-matched positive-evidence and counterfactual views using shared crop and rendering rules, and use a self-teacher to score student-generated tokens under both views with their prefixes held fixed. The prediction contrast between regions guides feedback placement and modulates GRPO advantages derived from task rewards, without reversing the signs of nonzero advantages. At inference, the model uses only the original image and question, without region annotations or additional visual tools.

Our main contributions are as follows:

- **Native-thinking post-training for fine-grained visual understanding.** We propose VC-RL, which combines visual feedback from local regions with sequence-level task rewards to optimize long-form thinking trajectories generated from full images and improve the use of fine-grained image information.
- **Visual feedback from matched regional views.** We construct teacher views from positive-evidence and counterfactual regions of equal size, using shared crop and rendering rules to control differences in visual conditions. Prediction contrast under the same student prefixes provides regional visual feedback for selective advantage modulation.
- **Visual performance gains across model scales.** VC-RL improves average accuracy over the 4B and 9B base models by 4.45 and 3.90 percentage points, respectively, across six fine-grained and high-resolution visual evaluations, with additional gains on multiple general multimodal tasks. These results indicate that post-training for fine-grained visual understanding also transfers to a broader range of visual tasks.

2 Related Work

2.1 Fine-Grained Visual Reasoning

High-resolution visual reasoning requires multimodal models to locate and use fine-grained evidence in complex images. Methods such as V*, DeepEyes, and DeepEyesV2 support reasoning through explicit visual search, cropping, or zooming, at the cost of additional visual processing during inference (Wu & Xie, 2024; Zheng et al., 2026; Hong et al., 2026). Other approaches transfer local perceptual capabilities to full-image models through training, enabling inference without additional visual operations (Wei et al., 2026; Yuan et al., 2026; Zhu et al., 2026; Li et al., 2026b). Visual on-policy distillation supports this transfer by providing teacher feedback at student-generated prefixes.

These methods differ in how they construct and select visual feedback. Vision-OPD uses an evidence-centered crop as privileged teacher context to provide fine-grained supervision on trajectories generated by a full-image student (Yuan et al., 2026). Beyond positive-view supervision, OPD-V introduces positive and negative visual conditions and selects tokens based on their probability differences (Aniri et al., 2026). Regional contrast also supports trajectory-level weighting: V-Zero contrasts relevant and irrelevant regions to estimate each student trajectory's dependence on visual evidence and adjusts distillation strength accordingly (Sun et al., 2026). At the token level, VAD compares teacher distributions with the relevant evidence present and removed to isolate corrections attributable to visual evidence (Zhang et al., 2026a). Together, these studies show that effective visual distillation depends not only on whether the teacher observes additional information, but also on whether its feedback is grounded in task-relevant visual evidence. Our work focuses on using regional visual feedback to optimize native thinking trajectories under sequence-level task rewards.

2.2 Reinforcement Learning and Self-Distillation

RLVR methods such as GRPO improve long-form reasoning by optimizing model-generated trajectories with verifiable outcome rewards (Shao et al., 2024; DeepSeek-AI, 2025). However, a single trajectory-level advantage is broadcast to every token, providing limited credit assignment to intermediate reasoning steps. OPD instead provides dense token-level feedback on student-sampled prefixes (Agarwal et al., 2024), while privileged-context OPSD creates an informational advantage for a teacher derived from the same model (Zhao et al., 2026).

Outcome rewards and privileged distillation provide complementary supervision, but teacher feedback does not necessarily improve long-form reasoning, as recent studies of textual thinking models demonstrate (Kaur et al., 2026). Recent work combines outcome rewards with on-policy self-distillation. SRPO routes samples between GRPO and self-distillation according to correctness and the availability of teacher feedback (Li et al., 2026a). RLSD uses self-distillation signals to adjust token-level update magnitudes while deriving update directions from environmental rewards (Yang et al., 2026). RLCSD contrasts privileged textual contexts and combines token selection, bounded sign-preserving advantage modulation, and separately normalized loss branches (Pan et al., 2026). Our work addresses the construction of teacher conditions for fine-grained visual tasks, where local views change both evidence visibility and surrounding visual context. We construct size-matched positive-evidence and counterfactual views under shared rendering rules, and use their prediction contrast to guide native visual thinking with sequence-level task rewards.

3 Method

Figure 2 summarizes VC-RL, comprising paired visual evidence construction, contrast-guided token selection, and verifier-anchored policy optimization.

3.1 Task Setup and Visual Evidence Construction

Each training example comprises a high-resolution image I, a question q, a verifiable answer $a^\star$, and a region annotation b^+ covering the relevant evidence. The student observes the full image I^{s} with this region marked by a red box and receives a corresponding instruction to attend to the indicated region. For each input $x = (I^{\mathrm{s}}, q)$, the rollout policy samples a group of G trajectories containing

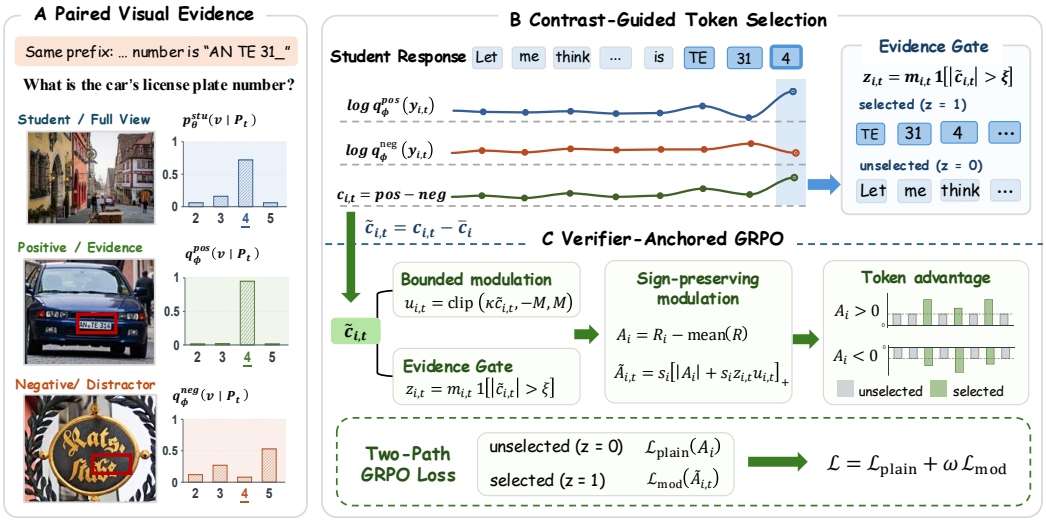


Figure 2: **Overview of VC-RL for native visual thinking.** (A) The student generates trajectories from full images; an EMA self-teacher scores them under matched positive-evidence and counterfactual views. (B) Response-centered teacher log-probability differences guide token selection. (C) Selected-token advantages receive bounded modulation without reversing nonzero advantage signs. Selected and unselected token losses are averaged separately within each micro-batch and combined with weight ω. Here, $s_i = \mathrm{sign}(A_i)$ and $[x]_+ = \max(x, 0)$.

both the thinking process and the final answer:

$$y_i = (y_{i,1}, \ldots, y_{i,T_i}) \sim p_{\theta_{\text{old}}}(\cdot \mid x), \qquad i = 1, \ldots, G. \tag{1}$$

Here, $p_{\theta_{\text{old}}}$ denotes the rollout policy, and q includes the question and the accompanying region-focus instruction.

As illustrated in Figure 2(A), we construct two teacher views from the same image. The positive evidence view I^+ highlights the annotated region b^+, whereas the counterfactual view I^- highlights another salient region b^- that does not overlap it. The two regions have equal width and height, and the expanded counterfactual crop excludes the annotated target region. Both views use the same crop scale, bounding-box rendering rules, and distance-dependent mosaic transformation to reduce variation from visual preprocessing.

A self-teacher q_ϕ, initialized from the same base model as the student and updated through an exponential moving average (EMA) of the student parameters, scores student trajectories under both visual conditions. For each generated token, we hold the question, textual instructions, and student-generated prefix fixed and use teacher forcing to compute its conditional log-probability under I^+ and I^-. All teacher scores and derived signals are detached from gradient computation.

The added region annotations, region-focus instructions, and auxiliary teacher views are used only during post-training. At inference, the student answers directly from the original benchmark image and question, without added region annotations or additional visual tools.

3.2 Regional Contrast and Token Selection

For each student-generated token, we define the log-likelihood difference between the two teacher views:

$$c_{i,t} = \log q_\phi(y_{i,t} \mid I^+, q, y_{i,<t}) - \log q_\phi(y_{i,t} \mid I^-, q, y_{i,<t}). \tag{2}$$

This contrast measures the relative support for the generated token under the two regional views. We use its response-centered value to guide token selection and advantage modulation.

To compare feedback across positions within each response, we center the raw contrasts. Let $m_{i,t} \in \{0, 1\}$ denote the valid response-token mask, excluding prompt and padding positions. We compute

the mean over all valid thinking and final-answer tokens and center their contrasts as follows:

$$\bar{c}_i = \frac{\sum_t m_{i,t} c_{i,t}}{\sum_t m_{i,t}}, \qquad \widetilde{c}_{i,t} = c_{i,t} - \bar{c}_i. \tag{3}$$

The centered signal represents each token's deviation from the response mean, without standard-deviation scaling.

As shown in Figure 2(B), we select positions whose deviation magnitude exceeds a fixed threshold $\xi > 0$:

$$z_{i,t} = m_{i,t}\, \mathbf{1}[|\widetilde{c}_{i,t}| > \xi]\,. \tag{4}$$

3.3 Verifier-Anchored Policy Optimization

Let R_i denote the sequence-level reward for trajectory y_i, including verifier feedback and reward-shaping terms defined in Appendix A.2. We compute sequence advantages by subtracting the group mean, without group standard-deviation normalization:

$$A_i = R_i - \frac{1}{G}\sum_{j=1}^{G} R_j, \tag{5}$$

We convert the signed centered contrast into a bounded token-level adjustment to the sequence advantage:

$$u_{i,t} = \mathrm{clip}(\kappa \widetilde{c}_{i,t}, -M, M)\,, \qquad u_{i,t} \in [-M, M]. \tag{6}$$

Here, $\kappa > 0$ controls the modulation slope and $M > 0$ bounds its magnitude. As illustrated in Figure 2(C), we adjust selected-token advantages and clamp them to prevent sign reversals of nonzero sequence advantages; unselected positions retain their original advantages:

$$\widetilde{A}_{i,t} = \begin{cases} \max(A_i + u_{i,t}, 0), & z_{i,t} = 1,\ A_i > 0, \\ \min(A_i + u_{i,t}, 0), & z_{i,t} = 1,\ A_i < 0, \\ 0, & z_{i,t} = 1,\ A_i = 0, \\ A_i, & z_{i,t} = 0. \end{cases} \tag{7}$$

Let $\mathcal{T}_0$ and $\mathcal{T}_1$ be the unselected and selected valid-token sets within an optimization micro-batch. We average the two sets separately:

$$\mathcal{L}_{\mathrm{plain}} = \frac{1}{|\mathcal{T}_0|}\sum_{(i,t)\in\mathcal{T}_0} \ell_{i,t}^{\mathrm{PPO}}(A_i), \qquad \mathcal{L}_{\mathrm{mod}} = \frac{1}{|\mathcal{T}_1|}\sum_{(i,t)\in\mathcal{T}_1} \ell_{i,t}^{\mathrm{PPO}}(\widetilde{A}_{i,t}), \tag{8}$$

and combine them as

$$\mathcal{L} = \mathcal{L}_{\mathrm{plain}} + \omega \mathcal{L}_{\mathrm{mod}} \tag{9}$$

where $\ell_{i,t}^{\mathrm{PPO}}$ is the standard clipped policy loss. Each branch averages over its valid tokens within the optimization micro-batch, with empty branches contributing zero; ω controls the relative weight of the modulated branch.

4 Experiments

4.1 Experimental Setup

Models and training data. We evaluate our method on Qwen3.5 (Qwen Team, 2026) at the 4B and 9B scales. Our post-training experiments use 6,159 high-resolution visual reasoning examples derived from Vision-OPD (Yuan et al., 2026). Following Vision-OPD and ZwZ (Yuan et al., 2026; Wei et al., 2026), we retain region markings and spatial instructions to disambiguate questions generated from local regions. GRPO, SDPO, Uniform OPD, and VC-RL share the same training examples and full-image student inputs, including these markings and instructions. Evaluation uses the original benchmark images and questions without added region guidance or auxiliary teacher views.

Table 1: Accuracy (%) on six visual benchmarks and their unweighted mean (Avg.). Bold marks the best score within each Qwen3.5 model size.

Model	Size	V*	Zoom	HR-Bench		MME-RealWorld		Avg.
				4K	8K	EN	CN	
Closed-source models								
GPT-5.2	–	79.06	50.89	81.12	78.38	72.60	68.80	71.81
GPT-5.4	–	76.96	52.66	84.00	77.88	74.20	70.93	72.77
Gemini-3.1-Pro	–	87.96	61.18	89.63	86.88	76.53	73.31	79.25
Gemini-3.5-Flash	–	89.01	61.42	89.12	86.62	75.31	73.97	79.24
Open-source models								
DeepEyes	7B	85.86	46.51	75.13	72.63	64.10	64.09	68.05
Thyme	7B	82.20	45.09	77.00	72.00	64.80	64.59	67.61
DeepEyesV2	7B	81.68	44.97	77.88	73.75	64.90	65.07	68.04
MiMo-VL-RL	7B	83.25	45.68	73.50	69.38	62.73	55.89	65.07
ZwZ	8B	87.96	56.69	83.63	81.75	66.57	68.09	74.12
GLM-4.6V	106B	86.91	50.06	82.13	78.88	65.57	65.62	71.53
Qwen3.5	397B	87.96	57.16	89.38	85.50	74.82	69.82	77.44
Kimi-K2.6	1T	88.48	53.14	81.88	78.00	69.22	66.13	72.81
Qwen3.5 models (native-thinking evaluation)								
Qwen3.5 Base	4B	86.39	52.78	85.25	81.25	73.43	69.46	74.76
+ Vision-OPD	4B	88.48	57.87	80.50	77.00	70.88	68.38	73.85
+ OPD-V	4B	89.53	59.05	83.00	78.62	71.90	69.70	75.30
+ Answer-hint control	4B	88.48	58.46	85.75	81.88	73.18	71.61	76.56
+ Ours	4B	**90.58**	**60.00**	**90.62**	**85.12**	**75.48**	**73.48**	**79.21**
Qwen3.5 Base	9B	89.53	57.04	88.12	85.12	74.18	70.56	77.43
+ Vision-OPD	9B	92.67	58.93	84.62	83.00	72.78	70.63	77.11
+ OPD-V	9B	93.19	60.71	86.25	84.00	74.74	71.30	78.36
+ Answer-hint control	9B	89.01	58.11	87.00	84.25	73.07	72.30	77.29
+ Ours	9B	**94.24**	**63.08**	**91.25**	**89.12**	**76.26**	**74.01**	**81.33**

Notes. All Qwen3.5-based models use native-thinking inference with a common generation configuration. Vision-OPD, OPD-V, and the answer-hint control are trained with native thinking enabled. The answer-hint control replaces regional teacher views with textual answer hints while retaining VC-RL's optimization settings.

Baselines. For broad comparison, we include the closed-source GPT-5.2 (OpenAI, 2025), GPT-5.4 (OpenAI, 2026), Gemini-3.1-Pro (Google DeepMind, 2026a), and Gemini-3.5-Flash (Google DeepMind, 2026b) models, together with the open-source DeepEyes (Zheng et al., 2026), Thyme (Zhang et al., 2026b), DeepEyesV2 (Hong et al., 2026), MiMo-VL-RL (Xiaomi LLM-Core Team, 2025), ZwZ (Wei et al., 2026), GLM-4.6V (GLM-V Team, 2025), and Kimi-K2.6 (Moonshot AI, 2026) models. On the Qwen3.5-4B and Qwen3.5-9B backbones, we compare against the unmodified base model and the Vision-OPD (Yuan et al., 2026) and OPD-V (Aniri et al., 2026) baselines, which we train with native thinking enabled. We also evaluate an answer-hint control at both scales to compare textual and regional teacher conditions under the same optimization settings. All Qwen3.5-based models are evaluated in native-thinking mode. The 9B implementations and training settings for these controls, GRPO, SDPO, Uniform OPD, and VC-RL are detailed in Appendix B.

Benchmarks and aggregate metric. We evaluate fine-grained and high-resolution visual perception with V* Bench (Wu & Xie, 2024), ZoomBench (Wei et al., 2026), HR-Bench (Wang et al., 2025) at 4K and 8K resolutions, and the English and Chinese splits of MME-RealWorld (Zhang et al., 2025). Together, these benchmarks cover small-object recognition, local evidence retrieval, high-resolution understanding, and reasoning in real-world scenes. We report accuracy on each benchmark and use their unweighted mean as the aggregate metric.

Table 2: **Training objectives and ablations on Qwen3.5-9B.** T^+: positive teacher view; $T^\pm$: paired positive/counterfactual views. Reweight.: separate normalization of selected/unselected branches. Avg6: mean accuracy (%) over six benchmarks. See Appendix D.4.

Variant	GRPO	Teacher	Distill.	Reweight.	Centering	Mod.	Scope	Avg6
GRPO	✓	–	–	–	–	–	–	77.67
SDPO	–	T^+	✓	–	–	–	–	78.87
Uniform OPD	✓	T^+	✓	–	–	–	–	78.84
RLCSD (visual)	✓	$T^\pm$	–	✓	–	tanh	All	80.09
Gate-only	✓	$T^\pm$	–	✓	✓	–	–	72.41
SeqMean	✓	$T^\pm$	–	–	✓	Clip	All	79.70
Dense Mod.	✓	$T^\pm$	–	–	✓	Clip	All	78.72
Shuffled Sel.	✓	$T^\pm$	–	✓	✓	Clip	All	77.52
Think-only	✓	$T^\pm$	–	✓	✓	Clip	Thinking	80.98
Answer-only	✓	$T^\pm$	–	✓	✓	Clip	Answer	78.08
Ours	✓	$T^\pm$	–	✓	✓	Clip	All	**81.33**

The shared generation settings, answer-scoring policy, and failure handling are specified in Appendix C.

Training configuration. Unless otherwise specified, our method samples eight on-policy thinking trajectories per prompt and performs full-parameter training for 120 optimization steps. The 9B main result uses centered VC-RL with an evidence threshold of $\xi = 0.30$, matching the reference configuration in Section 4.3.1. The shared 4B and 9B training configuration and teacher update rule are provided in Appendix A.1.

4.2 Main Results

Overall performance. Table 1 shows that VC-RL improves performance on fine-grained and high-resolution visual tasks at both model scales. Qwen3.5-4B and Qwen3.5-9B achieve average accuracies of **79.21%** and **81.33%**, respectively, exceeding their base models by 4.45 and 3.90 percentage points, with gains across all six benchmarks. The trained 4B model also surpasses the 9B base model in average accuracy, indicating that targeted visual post-training can improve a smaller model's use of fine-grained information.

Comparison with visual distillation methods. Vision-OPD and OPD-V improve on V* and ZoomBench but underperform their respective base models on HR-Bench 4K and 8K at both scales. VC-RL improves across these tasks, outperforming both baselines on all six benchmarks. At 9B, its average accuracy exceeds Vision-OPD and OPD-V by 4.22 and 2.97 percentage points, respectively, with consistent gains from local detail recognition to high-resolution understanding.

Regional views versus answer hints. To examine the role of regional teacher conditions, the answer-hint control replaces them with correct and incorrect answer-option hints while retaining the same optimization and training settings. Regional visual feedback yields higher accuracy at both model scales, supporting the use of local image conditions to construct contrastive feedback for fine-grained visual tasks.

4.3 Analysis and Ablations

4.3.1 Training Objectives and Component Ablations

We analyze how visual feedback is incorporated and where it is applied on Qwen3.5-9B. Table 2 summarizes the objective components and mean accuracy across six benchmarks, examining the effects of advantage modulation, loss aggregation, and feedback coverage. Per-benchmark results are provided in Appendix D.4.

Table 3: **Visual contrast construction on Qwen3.5-9B.** Accuracy (%) on six benchmarks and their unweighted mean (Avg6). Full-image ref. uses a full-image reference; Unmatched size and Ours use regional references without and with size matching, respectively. Bold marks the best Avg6.

Variant	V*	Zoom	HR-4K	HR-8K	MME-EN	MME-CN	Avg6
Full-image ref.	87.43	57.99	80.50	77.75	71.27	64.26	73.20
Unmatched size	93.19	63.43	90.38	86.00	76.49	74.11	80.60
Ours	94.24	63.08	91.25	89.12	76.26	74.01	**81.33**

Visual feedback in policy optimization. Uniform OPD performs similarly to teacher-only SDPO, suggesting limited benefit from directly combining sequence-reward and distillation losses. VC-RL uses visual contrast to adjust policy advantages, integrating both signals at the token level. It outperforms RLCSD under identical regional views; separate ablations support response-level centering and clipped-linear modulation (Appendix D.2). *SeqMean* retains modulation gains under GRPO's loss aggregation, while separate branch normalization further improves performance (Appendix B.3). *Gate-only* retains gating and branch normalization but removes modulation, yielding no comparable gains. These results support combining selective weighting with visually guided advantage adjustments; reweighting selected positions alone is insufficient.

Selective visual feedback. We examine the role of token selection through *Dense Mod.* and *Shuffled Sel.* Extending modulation to all valid tokens, or preserving the selection count and modulation values while shuffling their corresponding positions, reduces performance. These results indicate that the effectiveness of visual feedback depends not only on its coverage but also on where it is applied. Selecting tokens according to regional contrast helps concentrate feedback at positions that are more valuable for training.

Visual feedback during thinking. *Think-only* performs close to the full method and outperforms *Answer-only* on all six benchmarks (Table 10). Modulating thinking tokens alone retains much of the full method's gains, whereas applying feedback only to the final answer does not achieve comparable performance. These results identify intermediate thinking positions as an important locus for regional visual feedback.

Results across training data seeds. Across three training data seeds, VC-RL achieves an Avg6 of 81.09 ± 0.22, compared with 78.15 ± 0.68 for GRPO and 79.98 ± 0.52 for visually adapted RLCSD (mean $\pm$ sample standard deviation). Full results and offline analyses of regional contrast are provided in Appendix D.

4.3.2 Visual Contrast Construction

Effective regional contrast requires a reference view that helps distinguish feedback from the target region from differences introduced by changes in visual conditions. We evaluate two controls on Qwen3.5-9B: *Full-image ref.* replaces the counterfactual view with the full image, while *Unmatched size* retains the negative region's location but changes its size. All other training and evaluation settings remain fixed; Table 3 reports the results.

Local negative views provide a more effective reference. Even without size matching, a local negative view substantially outperforms the full-image reference. This indicates that a comparison condition excluding the target region is more effective than directly contrasting a local view with the full image. The full image already contains the target evidence, so this comparison changes how the evidence is presented and its surrounding context. A local negative view instead allows teacher feedback to contrast how different regional contents support the same student trajectory.

Size matching further improves regional contrast. With a local negative view in place, matching region sizes further improves average performance. Equal region widths and heights, together with shared crop and rendering rules, provide more consistent spatial conditions for the two views,

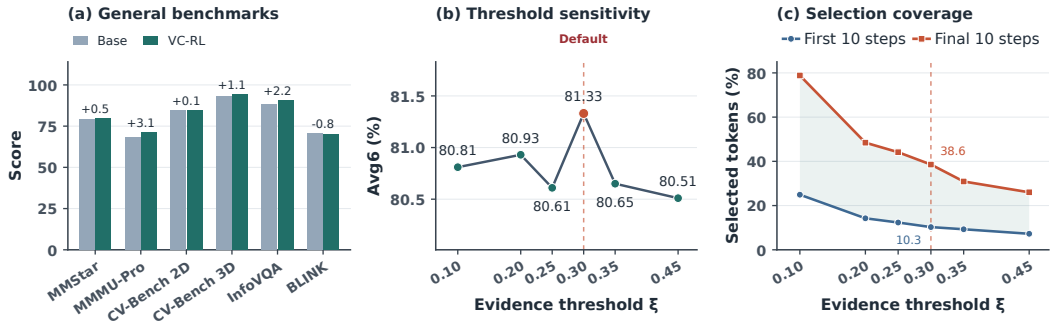


Figure 3: **Generalization and evidence-threshold sensitivity on Qwen3.5-9B.** (a) Performance on general multimodal benchmarks; annotations indicate gains over the base model. (b) Avg6 across evidence thresholds ξ. (c) Mean token selection rates over the first and last ten training steps; shading indicates their difference.

limiting variation due to region size in the comparison of teacher predictions. Together, these comparisons support the local negative reference, with size matching yielding the highest average accuracy among the evaluated constructions.

4.3.3 General Multimodal Evaluation

We evaluate the generalization of VC-RL beyond high-resolution visual tasks on MMStar (Chen et al., 2024), MMMU-Pro Vision (Yue et al., 2025), CV-Bench 2D and 3D (Tong et al., 2024), InfoVQA-val (Mathew et al., 2022), and BLINK (Fu et al., 2025). Figure 3(a) compares VC-RL with the Qwen3.5-9B base model. We report ANLS for InfoVQA and accuracy for the remaining benchmarks, all on a 0–100 scale. VC-RL outperforms the base model on five of the six benchmarks, with the largest gains on MMMU-Pro Vision and InfoVQA and a slight decline on BLINK. The improvements span multimodal reasoning, document question answering, and visual perception, demonstrating that visual contrast-guided post-training also benefits general multimodal tasks.

4.3.4 Effect of the Evidence Threshold

We vary the evidence threshold on Qwen3.5-9B to examine how feedback coverage affects performance, keeping all other training settings fixed. Higher thresholds restrict modulation to tokens whose contrast deviates further from the response mean.

Figure 3(b–c) shows substantial differences in token selection rates across thresholds, while Avg6 varies by at most 0.82 percentage points. Broader coverage does not consistently improve performance, and the sparser feedback at higher thresholds retains comparable accuracy. These results indicate that VC-RL does not require a narrowly tuned selection rate. At a fixed threshold, selection rates also differ between the first and last ten training steps: the gate imposes no fixed token quota, allowing feedback coverage to change with the contrast distribution over training.

We select the default threshold using a held-out validation set of 600 examples from the same task distribution, with no source-image overlap with the training or evaluation sets. The selected value, $\xi = 0.30$, is fixed across the reported benchmarks for the default 9B configuration.

5 Conclusion

We introduced VC-RL, which combines GRPO with visual on-policy self-distillation for native visual thinking. Contrast between matched regional views selectively modulates token-level advantages without reversing nonzero sequence-advantage signs. Experiments at two model scales show gains on fine-grained and high-resolution visual tasks, with further improvements on general multimodal benchmarks. Ablations highlight the importance of selecting feedback positions and applying regional supervision during thinking.

AI USE STATEMENT

We used generative AI tools for writing assistance and translation, literature search, and feedback on research ideas, experimental design, and result interpretation. The use of GPT-OSS-20B for answer mapping is described in Appendix C.2. The authors reviewed and revised the AI-assisted content and take responsibility for the final manuscript and reported results.

REPRODUCIBILITY STATEMENT

To support reproducibility, we describe the method in Section 3 and provide training configurations, reward definitions, and visual input construction in Appendix A. Baseline implementations and loss aggregation are described in Appendix B. Appendix C specifies the evaluation protocol and random-seed conventions, while Appendix D reports additional comparisons and multi-seed 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.

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. *CoRR*, abs/2608.05131, 2026. doi: 10.48550/arXiv.2608.05131. URL https://doi.org/10.48550/arXiv.2608.05131.

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.

DeepSeek-AI. DeepSeek-R1: Incentivizing reasoning capability in LLMs via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025. URL https://arxiv.org/abs/2501.12948.

Xingyu Fu, Yushi Hu, Bangzheng Li, Yu Feng, Haoyu Wang, Xudong Lin, Dan Roth, Noah A. Smith, Wei-Chiu Ma, and Ranjay Krishna. BLINK: Multimodal large language models can see but not perceive. In *Computer Vision – ECCV 2024*, volume 15081 of *Lecture Notes in Computer Science*, pp. 148–166. Springer Nature Switzerland, 2025. doi: 10.1007/978-3-031-73337-6_9. URL https://doi.org/10.1007/978-3-031-73337-6_9.

GLM-V Team. GLM-4.5V and GLM-4.1V-Thinking: Towards versatile multimodal reasoning with scalable reinforcement learning, 2025. URL https://arxiv.org/abs/2507.01006.

Google DeepMind. Gemini 3.1 Pro. https://deepmind.google/models/gemini/pro/, 2026a.

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

Jack Hong, Chenxiao Zhao, ChengLIn Zhu, Weiheng Lu, Guohai Xu, and XingYu. DeepEyesV2: Toward agentic multimodal model. In C. Vondrick, B. Hariharan, C. Raffel, L. Pinto, D. Yang, and A. Faust (eds.), *International Conference on Learning Representations*, volume 2026, pp. 114851–114872, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/file/badbef7d421640fb038399e9d7271b1c-Paper-Conference.pdf.

Simran Kaur, Narutatsu Ri, Yinghui He, Liam Fowl, and Sanjeev Arora. Rethinking on-policy self-distillation for thinking models. *arXiv preprint arXiv:2607.05184*, 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*, 2026a.

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*, 2026b.

Minesh Mathew, Viraj Bagal, Rubèn Tito, Dimosthenis Karatzas, Ernest Valveny, and CV Jawahar. Infographicvqa. In *2022 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 2582–2591. IEEE, 2022.

Moonshot AI. Kimi K2.6: Advancing open-source coding. https://www.kimi.com/en/blog/kimi-k2-6, 2026.

OpenAI. GPT-5.2 Model. https://developers.openai.com/api/docs/models/gpt-5.2, 2025.

OpenAI. GPT-5.4 Model. https://developers.openai.com/api/docs/models/gpt-5.4, 2026.

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. https://qwen.ai/blog?id=qwen3.5, February 2026.

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.

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.

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, 2024.

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 *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13084–13094. IEEE, 2024.

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

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.

Xiang Yue, Tianyu Zheng, Yuansheng Ni, Yubo Wang, Kai Zhang, Shengbang Tong, Yuxuan Sun, Botao Yu, Ge Zhang, Huan Sun, Yu Su, Wenhu Chen, and Graham Neubig. MMMU-Pro: A more robust multi-discipline multimodal understanding benchmark. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 15134–15186, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.736. URL https://aclanthology.org/2025.acl-long.736/.

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*, 2026a.

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. In *International Conference on Learning Representations*, 2026b.

YiFan Zhang, Huanyu Zhang, Haochen Tian, Chaoyou Fu, Shuangqing Zhang, Junfei Wu, Feng Li, Kun Wang, Qingsong Wen, Zhang Zhang, Liang Wang, and Rong Jin. MME-RealWorld: Could your multimodal llm challenge high-resolution real-world scenarios that are difficult for humans? In Y. Yue, A. Garg, N. Peng, F. Sha, and R. Yu (eds.), *International Conference on Learning Representations*, volume 2025, pp. 89655–89701, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/df29d63af05cb91d705cf06ba5945b9d-Paper-Conference.pdf.

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.

Ziwei Zheng, Minghao Yang, Jack Hong, Chenxiao Zhao, Guohai Xu, Le Yang, Chao Shen, and XingYu. DeepEyes: Incentivizing "thinking with images" via reinforcement learning. In C. Vondrick, B. Hariharan, C. Raffel, L. Pinto, D. Yang, and A. Faust (eds.), *International Conference on Learning Representations*, volume 2026, pp. 126775–126798, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/file/cdb347d7516d52a9280ea9d6708911f6-Paper-Conference.pdf.

Qihui Zhu, Yuchen Wang, Zijian Wen, Tao Zhang, Mengjie Zhang, Yang Liu, Shuangwu Chen, Siying Wu, Jian Yang, and Xiaofeng Jiang. RP-OPSD: Resolution-privileged on-policy self-distillation for multimodal large language models. *arXiv preprint arXiv:2607.24447*, 2026.

A Implementation Details

A.1 Training Configuration

Table 4 summarizes the shared default configuration for centered VC-RL on Qwen3.5-4B and Qwen3.5-9B. We perform full-parameter post-training with 24 prompts per training batch and eight sampled trajectories per prompt, yielding 192 trajectories per rollout batch. The learning rate follows a cosine schedule with 10 warmup steps and a minimum learning rate equal to 10% of the peak value.

Table 4: Shared default training configuration for centered VC-RL on Qwen3.5-4B and Qwen3.5-9B.

Parameter	Value
Evidence threshold ξ	0.30
Modulation slope κ	1.5
Modulation cap M	0.5
Modulated-branch weight ω	1.0
EMA student update rate α	0.01
Peak learning rate	2×10^{-6}
Minimum learning rate	2×10^{-7}
Learning-rate schedule	Cosine
Warmup steps	10
Training batch size	24 prompts
Rollouts per prompt	8
Training steps	120
Maximum response length	8,192 tokens

With the default $\kappa = 1.5$, $\xi = 0.30$, and $M = 0.5$, selected tokens with $0.30 < |\widetilde{c}_{i,t}| < 1/3$ receive unsaturated modulation; those with $|\widetilde{c}_{i,t}| \geq 1/3$ receive increments of $\pm M$ before advantage clamping.

The self-teacher is initialized from the same base checkpoint as the student. After each actor-update phase that performs a parameter update, we update the teacher parameters as

$$\phi \leftarrow (1-\alpha)\phi + \alpha\theta, \qquad \alpha = 0.01. \tag{10}$$

The same teacher is used for both visual conditions, and teacher scores and all derived modulation signals are detached from gradient computation.

A.2 Reward Definition and Answer Verification

At both model scales, the sequence reward combines answer correctness with completion, length, and repetition terms. Let h_i indicate that the response contains a closing thinking tag, v_i indicate that an answer can be extracted after the final closing tag, and a_i indicate that this valid answer matches the ground truth. We use

$$R_i = a_i + 0.10h_i - 0.75(1 - v_i) - P_{\text{len}}(L_i) - P_{\text{rep}}(d_i), \tag{11}$$

where L_i is the number of generated response tokens, including thinking and final-answer tokens. The two penalties are

$$\begin{aligned} P_{\text{len}}(L_i) &= 0.25 \ \text{clip}\left(\frac{L_i - 4096}{4096}, 0, 1\right), \\ P_{\text{rep}}(d_i) &= 0.30 \ \text{clip}\left(\frac{d_i - 0.15}{0.20}, 0, 1\right). \end{aligned} \tag{12}$$

For $L_i \geq 8$, let U_i be the number of distinct contiguous 8-grams in the response token sequence. We define $d_i = 1 - U_i/(L_i - 7)$; for shorter responses, $d_i = 0$. The length penalty starts above 4,096 tokens and reaches 0.25 at 8,192 tokens. The repetition penalty starts above a duplicate ratio of 0.15 and reaches 0.30 at 0.35. No additional penalty is applied solely because a response is truncated; its generated content is evaluated by the same rules.

A.3 Visual Input Construction

Paired regions. The student uses the full annotated image and its region-focus instruction. Teacher views are constructed from the corresponding original image and the positive pixel-coordinate box b^+. Coordinates are converted to integers and clipped to the image boundaries. For each example, we propose negative boxes with exactly the same width w and height h as b^+, fully contained within the source image. Candidate top-left positions comprise a 7×7 grid and 512 uniformly sampled positions. Duplicate candidates are removed. A candidate must be separated from b^+ along at least one axis by a margin of $\lceil 0.75w \rceil$ horizontally or $\lceil 0.75h \rceil$ vertically. This ensures that its $2.5\times$ expanded crop does not intersect the annotated positive box.

Negative-region selection. We rank eligible candidates using a grayscale thumbnail whose longest side is at most 512 pixels. For a candidate region b, the score is

$$s(b) = \frac{\sigma_b}{64} + \frac{g_b}{48} + 0.08\,\frac{\|\mathrm{center}(b) - \mathrm{center}(b^+)\|_2}{\sqrt{W^2 + H^2}}, \tag{13}$$

where σ_b is the standard deviation of grayscale intensities on the 0–255 scale, g_b is the mean sum of absolute horizontal and vertical intensity differences, and (W, H) is the source-image size. We select the highest-scoring candidate, with sampling and tie-breaking determined by fixed per-example seeds. Negative regions are selected automatically by this texture-based score, without human verification of their semantic content. The construction excludes 82 examples without an eligible negative box, retaining 6,159 examples.

Crop and distance-dependent mosaic. For either selected box, we expand its width and height by 2.5 about its center, convert the crop coordinates to integers, and clip them to the source-image boundaries. Pixels inside the box are retained before resizing and box rendering. Outside the box, let d be the Euclidean distance to the box and $d_{\max}$ the largest such distance in the crop. Pixels are assigned to five rings by $r = \min(\lfloor 5d/d_{\max} \rfloor, 4)$. For ring $r \in \{0, \ldots, 4\}$, we apply a pixelation factor $f_r = 2 \cdot 6^{r/4}$: the crop is bilinearly downsampled by f_r and resized back by nearest-neighbor interpolation, and the resulting pixels are copied into that ring.

For a crop of size (W_c, H_c), the candidate enlargement factor is

$$s_{\mathrm{up}} = \min\left\{\max\left(1, \frac{224}{\min(w, h)}\right), \frac{2048}{\max(W_c, H_c)}\right\}. \tag{14}$$

We apply bicubic resizing only when $s_{\mathrm{up}} > 1$; otherwise, the crop retains its dimensions. The 2,048-pixel bound applies to the longest side during enlargement. A red outline is drawn outward around the transformed box, with width $\max(3, \lfloor 0.004 \max(W_r, H_r) \rfloor)$ pixels for rendered dimensions (W_r, H_r). Finally, the negative view is bicubically resized to the positive view's exact output dimensions when necessary. This handles differences in crop dimensions caused by image-boundary clipping. Both views use the same rendering parameters and retain the same question, textual instructions, and student-generated prefix during teacher scoring.

A.4 Gate-only Control

The Gate-only control uses the same response-centered evidence gate, evidence threshold $\xi = 0.30$, branch weight $\omega = 1$, EMA teacher, training data, training budget, and optimization configuration as centered VC-RL. It also retains the two separately normalized branches in Eqs. (8)–(9). Gate-only sets $u_{i,t} = 0$, so selected tokens use the original sequence advantage A_i.

B Baseline Implementations

B.1 Shared Training Protocol

The 9B comparisons among GRPO, SDPO, Uniform OPD, and VC-RL use Qwen3.5-9B and the same 6,159 training examples with identical full-image student inputs. Table 5 lists the optimization and rollout settings shared across these methods, with objective-specific components applied where relevant. We train these methods in our codebase with native thinking enabled and an 8,192-token

rollout response limit. Teacher-based objectives score student-generated prefixes. The reward-driven objectives use the complete reward in Appendix A.2, including correctness, thinking completion, answer validity, length, and repetition terms. Teacher-only SDPO does not include this reward in its optimization objective.

Table 5: Shared training and rollout settings for GRPO, SDPO, Uniform OPD, and VC-RL at the 9B scale. Objective-specific settings apply only to methods that use the corresponding objective.

Setting	Value
Global prompt batch size	24
Rollouts per prompt G	8
Peak learning rate	2×10^{-6}
Learning-rate schedule	Cosine; minimum ratio 0.1
Warmup / training steps	10 / 120
Rollout temperature / top-p	0.9 / 0.95
Rollout repetition penalty	1.03
Maximum prompt / response length	8,192 / 8,192 tokens
PPO lower / upper clipping offset	0.2 / 0.3
Additional KL regularization	Disabled
Group standard-deviation normalization	Disabled
Maximum reward length penalty	0.25

B.2 Training Objectives and Method-Specific Settings

GRPO. The reinforcement-learning baseline uses the group-mean-centered sequence advantage in Eq. 5, without division by the group standard deviation. It optimizes the clipped policy objective with no teacher feedback or distillation term.

SDPO (teacher-only). The student is distilled from a self-teacher conditioned on the positive evidence view. The reverse-KL direction is $D_{\mathrm{KL}}(p_\theta \| q_\phi^+)$ at student-generated prefixes, with no forward-KL mixture. Distillation covers all valid response positions, including thinking and final-answer tokens. The objective is $\mathcal{L}_{\mathrm{SDPO}} = \mathcal{L}_{\mathrm{RKL}}$, with coefficient 1.0 and no verifier-driven policy term. The teacher is initialized from the base model and updated using the EMA rule in Eq. 10, with student update rate 0.01.

Uniform OPD. This baseline adds the same reverse-KL distillation objective to GRPO:

$$\mathcal{L}_{\mathrm{Uniform}} = \mathcal{L}_{\mathrm{GRPO}} + 0.6\,\mathcal{L}_{\mathrm{RKL}}. \tag{15}$$

The teacher condition, EMA update rate, and distillation direction match SDPO. Distillation is applied uniformly to all valid thinking and final-answer tokens, without evidence gating.

Vision-OPD and OPD-V. We train Vision-OPD (Yuan et al., 2026) and OPD-V (Aniri et al., 2026) with native thinking enabled using their respective distillation objectives. Vision-OPD uses JSD distillation with coefficient 1.0, cropped evidence views, token-mean loss aggregation, and a constant learning rate of 2×10^{-6}. OPD-V uses positive and negative visual conditions with a probability-difference selection threshold of zero. For both methods, Table 1 reports results with a teacher EMA student update rate of 0.01 at both model scales, matching VC-RL.

Answer-hint control. Inspired by RLCSD's contrastive textual conditioning (Pan et al., 2026), this control uses correct and incorrect answer-option hints to study the source of teacher feedback. Both teacher branches see the full student image and receive a shared hint template containing either the correct option or a deterministically selected incorrect option. The centered contrast gate, bounded modulation, two-branch loss, and training settings match VC-RL. This is a teacher-condition control within VC-RL, rather than a reproduction of the full RLCSD method.

RLCSD (visual adaptation). This baseline uses the same positive-evidence and counterfactual regional views as VC-RL, with RLCSD's contrast processing, token selection, bounded advantage modulation, and loss aggregation rules (Pan et al., 2026). Appendix D.2 reports the seed-42 comparison; results across three training seeds are provided in Table 7.

VC-RL. Our method uses Eqs. 5–9, response-level contrast centering, $\xi = 0.30$, $\kappa = 1.5$, $M = 0.5$, and $\omega = 1.0$. Both teacher conditions use the same EMA teacher with student update rate 0.01. No explicit reverse-KL or forward-KL distillation loss is added to the policy objective.

B.3 Loss Aggregation

GRPO, SDPO, and Uniform OPD average valid-token losses within each response and then across responses. VC-RL, the answer-hint control, and Gate-only instead average selected and unselected valid tokens separately within each optimization micro-batch (Eq. 8); an empty branch contributes zero. SeqMean retains VC-RL's contrast, token-selection gate, and advantage modulation but uses the response-mean aggregation of GRPO, without separately normalizing the selected and unselected branches.

For the default 9B run, we use AdamW ($\beta_1 = 0.9$, $\beta_2 = 0.999$, weight decay 0.01), gradient-norm clipping at 1.0, and actor seed 42. Dynamic batching uses a budget of 16,384 input tokens per device. Micro-batch losses are weighted by their response counts before gradient accumulation and FSDP synchronization.

C Evaluation Protocol

C.1 Inputs and Thinking-Mode Generation

All Qwen3.5-based models, including the base models, adapted baselines, and ablations, use the same evaluation pipeline on the six main benchmarks: V* Bench, ZoomBench, HR-Bench 4K and 8K, and MME-RealWorld EN and CN. We evaluate the complete official evaluation splits and compute Avg6 as the unweighted mean of their accuracies. The task-specific metrics for the general multimodal evaluation are described in Section 4.3.3.

Unless otherwise stated, single-run results from our post-training experiments use the default training data seed 42. Table 10 reports these single-run results, whereas Table 7 reports the mean and sample standard deviation across three training data seeds. The request-level evaluation sampling seed is fixed at 42 for all runs (Table 6).

We supply the original benchmark image, official question text, and a common multiple-choice instruction, without added region annotations or teacher views. Images are submitted without external resizing, except for V* images exceeding the 20 MB API payload limit: these are iteratively downsampled by a factor of 0.75 in each spatial dimension using Lanczos resampling until they satisfy the limit. We then apply the model's image processor and the official Qwen3.5 chat template with native thinking enabled.

Table 6: Shared evaluation generation settings for Qwen3.5-based models on the six main benchmarks.

Parameter	Value
Inference engine	vLLM
Maximum generated tokens	32,768
Temperature	1.0
Top-p / top-k / min-p	0.95 / 20 / 0
Presence penalty	1.5
Repetition penalty	1.0
Generations per question	1
Request-level sampling seed	42
Thinking mode	Enabled

C.2 Answer Extraction and Scoring

We use the same answer-extraction procedure for all Qwen3.5-based models on the six main benchmarks. The procedure first searches for an explicit option in `<answer>` tags, then applies deterministic option-extraction rules from VLMEvalKit. If these stages do not resolve the answer, a language-model mapper operating at temperature zero maps the response to an answer option. The recorded 9B evaluations use GPT-OSS-20B with a maximum of 16,384 generated tokens for this mapping stage. Accuracy is determined by comparing the extracted option with the ground truth.

Empty responses and responses from which no answer can be recovered are counted as incorrect, without random guessing or additional model generations. A truncated response remains eligible for scoring when an answer can be extracted, regardless of whether the thinking segment is closed.

Failed inference requests are retried, and scoring proceeds only after predictions are available for the complete evaluation split. Judge requests that are truncated are retried; unresolved mappings are counted as incorrect. All compared models use this same scoring policy.

D Additional Empirical Analyses

D.1 Results Across Training Seeds

We repeat training three times for each method with different training data seeds and evaluate all runs using the same protocol. Table 7 reports the mean and sample standard deviation. VC-RL achieves an Avg6 of 81.09 ± 0.22, compared with 78.15 ± 0.68 for GRPO. Removing response-level centering reduces the mean Avg6 to 80.40 ± 0.42, with lower scores in all three corresponding runs.

Gate-only exhibits greater variation across training seeds. Shuffled selection and unmatched region sizes both yield lower Avg6 than VC-RL in all three corresponding runs; the mean gain from size matching is 0.38 percentage points.

Table 7: Performance across three training runs on Qwen3.5-9B (mean $\pm$ sample standard deviation). Avg6 is the unweighted mean over the six benchmarks.

Method	V*	Zoom	HR 4K	HR 8K	MME EN	MME CN	Avg6
GRPO	90.58 $\pm$0.00	58.03 $\pm$2.01	88.46 $\pm$2.32	84.46 $\pm$1.06	74.62 $\pm$0.22	72.72 $\pm$0.23	78.15 $\pm$0.68
w/o Centering	92.32 $\pm$1.68	62.61 $\pm$1.14	90.04 $\pm$0.85	87.00 $\pm$0.12	76.29 $\pm$0.26	74.12 $\pm$0.33	80.40 $\pm$0.42
Gate-only	89.53 $\pm$1.05	55.46 $\pm$3.02	87.87 $\pm$2.19	81.71 $\pm$3.63	71.73 $\pm$4.17	68.72 $\pm$4.73	75.84 $\pm$3.07
Shuffled Sel.	89.53 $\pm$0.91	56.49 $\pm$1.62	88.83 $\pm$1.69	84.79 $\pm$0.89	74.79 $\pm$0.82	72.34 $\pm$0.75	77.79 $\pm$0.88
Unmatched size	93.02 $\pm$0.30	64.06 $\pm$0.55	89.92 $\pm$0.41	86.63 $\pm$0.55	76.52 $\pm$0.08	74.12 $\pm$0.11	80.71 $\pm$0.10
RLCSD (visual)	93.37 $\pm$1.09	61.22 $\pm$0.89	89.33 $\pm$1.02	86.50 $\pm$1.19	75.72 $\pm$0.91	73.73 $\pm$1.18	79.98 $\pm$0.52
VC-RL	93.89 $\pm$1.09	63.91 $\pm$0.72	90.50 $\pm$0.66	87.79 $\pm$1.53	76.37 $\pm$0.10	74.11 $\pm$0.35	81.09 $\pm$0.22

D.2 Comparison with Visually Adapted RLCSD

To compare the optimization designs under the same regional teacher conditions, we adapt RLCSD (Pan et al., 2026) to use the positive-evidence and counterfactual views constructed for VC-RL. This comparison holds the paired visual inputs fixed and examines how their contrast is incorporated into policy optimization.

RLCSD applies a bounded hyperbolic-tangent transformation to the raw contrast and selects tokens by the transformed magnitude. We examine two separate modifications to this baseline: + *Our modulation* replaces the hyperbolic-tangent mapping with our clipped-linear function without centering, while + *Centering* introduces response-level centering before selection and modulation while retaining the hyperbolic-tangent function. Unlike the *w/o Centering* variant in Table 7, which removes centering from VC-RL, + *Our modulation* replaces only the modulation function in visually adapted RLCSD, retaining its remaining optimization rules. Full VC-RL uses both response-level centering and clipped-linear modulation, with the optimization settings described in Appendix B. Table 8 reports all four variants with the default training data seed 42; Table 7 separately reports statistics across three training seeds.

Table 8: **Comparison with visually adapted RLCSD on Qwen3.5-9B.** All results are single-run accuracies (%) with the default training data seed 42 and the same paired regional views. Avg6 is the unweighted mean over six benchmarks; bold marks the best Avg6.

Method	**V***	**Zoom**	HR 4K	HR 8K	MME EN	MME CN	**Avg6**
RLCSD (visual)	93.72	62.13	88.62	85.12	76.53	74.43	80.09
+ Our modulation	92.15	62.13	90.38	86.50	76.61	74.07	80.31
+ Centering	93.19	61.78	91.00	87.00	76.35	74.13	80.58
VC-RL	94.24	63.08	91.25	89.12	76.26	74.01	**81.33**

Both modifications improve Avg6 over visually adapted RLCSD, with the larger increase obtained from centering in this run. Full VC-RL achieves the highest Avg6 of 81.33.

D.3 Offline Analysis of Regional Contrast

We examine how centering changes regional contrast using a fixed Qwen3.5-9B VC-RL checkpoint to score recorded training rollouts under the positive and counterfactual views. Each sampled group contains responses to the same prompt from the same training step, with at least two correct and two incorrect retained responses. Truncated responses, responses without closed thinking or an extractable answer, and responses with a duplicate 8-gram ratio above 0.30 are excluded. We score up to the first 2,048 tokens of each response and compute the contrast mean over this scored segment, using $\xi = 0.30$ for both raw and centered selection.

Table 9: Offline regional-contrast diagnostics on Qwen3.5-9B. Distribution and coverage statistics use 303 responses from 40 groups; sign changes are measured on a larger sample of 200 groups.

Statistic	Value
Mean / standard deviation of the segment mean contrast	$0.162 \,/\, 0.327$
Responses with absolute mean contrast above ξ	33.7%
Mean selection rate before centering	38.33%
Mean selection rate after centering	57.73%
Sign-change rate at selected positions after centering	12.1%

Selection rates are averaged over responses. Sign-change rates are computed within each response over positions selected by the centered gate, then averaged within and across scoring shards; responses with no selected positions are omitted from this statistic, and raw contrasts with magnitude below 10^{-8} are treated as unchanged.

Table 9 shows that the variation in mean contrast is comparable to the selection threshold, and centering increases selection coverage in this sample. Subtracting the segment mean therefore materially changes where the fixed threshold applies. At selected positions, the centered and raw contrasts retain the same sign in most cases, while the observed sign changes quantify how centering redistributes feedback relative to the mean support within each scored segment.

D.4 Per-Benchmark Component Ablation Results

Table 10 provides the complete benchmark results for the training objectives and component ablations summarized in Table 2. All variants use Qwen3.5-9B and the training and evaluation settings described in Appendices B and C.

In Table 2, T^+ and $T^\pm$ denote positive and paired positive/counterfactual teacher views. Distill. indicates explicit distillation. Reweight. indicates separate normalization of selected and unselected token branches. SeqMean retains the token-selection gate but uses GRPO's response-mean aggregation, so it is marked "–" in this column. Centering indicates response-level contrast centering. Mod. specifies the bounded signed modulation function: $\tanh$ denotes the hyperbolic-tangent mapping, and Clip denotes the clipped-linear mapping in Eq. 6. Scope specifies the segments eligible for modulation; All includes thinking and final-answer tokens. Avg6 is the unweighted mean accuracy over the six benchmarks, and bold marks the best score.

RLCSD (visual) uses the default training data seed 42, matching the run reported in Table 8. Further details of the contrast processing and optimization rules are provided in Appendices D.2 and B.

Table 10: **Per-benchmark results for training objectives and component ablations on Qwen3.5-9B.** Rows correspond to Table 2. We report accuracy (%) on V* Bench, ZoomBench, HR-Bench 4K and 8K, and MME-RealWorld EN and CN. Avg6 is their unweighted mean. Bold marks the best score in each column.

Variant	V*	Zoom	HR 4K	HR 8K	MME EN	MME CN	Avg6
GRPO	90.58	58.82	86.25	83.25	74.38	72.74	77.67
SDPO	93.19	60.00	88.00	84.38	74.60	73.06	78.87
Uniform OPD	93.19	59.88	87.12	86.00	74.93	71.95	78.84
RLCSD (visual)	93.72	62.13	88.62	85.12	76.53	**74.43**	80.09
Gate-only	88.48	52.07	85.62	77.75	67.00	63.55	72.41
SeqMean	92.67	60.71	88.88	87.00	75.30	73.62	79.70
Dense Mod.	92.67	58.34	88.50	85.00	73.57	74.24	78.72
Shuffled Sel.	90.58	55.74	88.62	84.00	74.30	71.86	77.52
Think-only	**94.76**	62.60	90.00	87.88	**76.54**	74.09	80.98
Answer-only	90.58	56.92	89.75	84.00	75.30	71.91	78.08
Ours	94.24	**63.08**	**91.25**	**89.12**	76.26	74.01	**81.33**