REGION-LEVEL CREDIT ASSIGNMENT FROM ON-POLICY ROLLOUTS IN MULTIMODAL REINFORCEMENT POST-TRAINING

Anonymous authors
Paper under double-blind review

ABSTRACT

Multimodal reinforcement post-training can improve a response's reward without resolving the visual mistakes within it. A shared outcome advantage can reinforce incorrect regional selections alongside correct ones, while omitted objects receive no explicit target in the sampled response. This leaves a gap between optimizing response quality and learning accurate, complete visual grounding. We present MoVE (**M**odel's **o**wn **V**isual **E**rrors), a framework that turns **the policy's own regional errors into corrective supervision**. MoVE contrasts corrected and error-preserving regional conditions to refine credit within a sampled response, and distills corrected targets to supervise objects absent from it. These complementary signals address both incorrect selections and omissions, tying training feedback to specific, verifiable regional mistakes. MoVE provides **a unified framework for region-aware multimodal post-training**, extending GRPO, DAPO, and OPD with targeted visual correction. Inference still uses only the original image and query. Across five grounding benchmarks, MoVE improves average Micro-F1 in matched GRPO and DAPO comparisons at both 2B and 4B, including an increase from 60.25 to 61.37 over DAPO with Qwen3-VL-2B. Corrected-region training benefits three model families, while the extension to visual question answering improves mean accuracy by 8.7% relative to the base models across four fine-grained benchmarks and two model scales.

1 INTRODUCTION

Multimodal large language models can answer questions and reason over entire images (Bai et al., 2025; Li et al., 2024; Chen et al., 2024; Deitke et al., 2025). For region-sensitive tasks, success depends on finding the right objects and retaining the local details required by the query. In visual grounding, these regions are the prediction itself, while in fine-grained visual question answering (VQA), they provide the information needed to answer. Multimodal post-training therefore needs feedback that improves both the correctness and completeness of the model's regional predictions.

A single response can nevertheless combine correct localizations, distracting regions, and missed objects. We call these partially correct responses *mixed-evidence rollouts*, where visual evidence refers to the regions selected to support the prediction. The two rollouts in Figure 1 attain the **same F1 score but have different error compositions**, with different counts of true positives (TP), false positives (FP), and false negatives (FN). In our analysis of 500 on-policy rollout groups with eight responses each, 490 groups contain distinct responses with the same F1 score. The illustrated responses require different corrections, yet their common score does not identify which regions to retain, remove, or recover.

This ambiguity has consequences for learning. GRPO (Shao et al., 2024) derives relative advantages from response rewards, while DAPO (Yu et al., 2025) improves clipping, sampling, and optimization stability. Visual-RFT (Liu et al., 2025), VLM-R1 (Shen et al., 2025), and Perception-R1 (Xiao et al., 2026) bring verifiable outcome supervision to visual question answering, grounding, and multimodal reasoning. In the group-relative objectives considered here, one outcome advantage weights all sampled tokens in a response. Consequently, **incorrect regions can be reinforced alongside correct ones**, and a negative advantage can penalize tokens that describe valid regions. Missed objects pose

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.

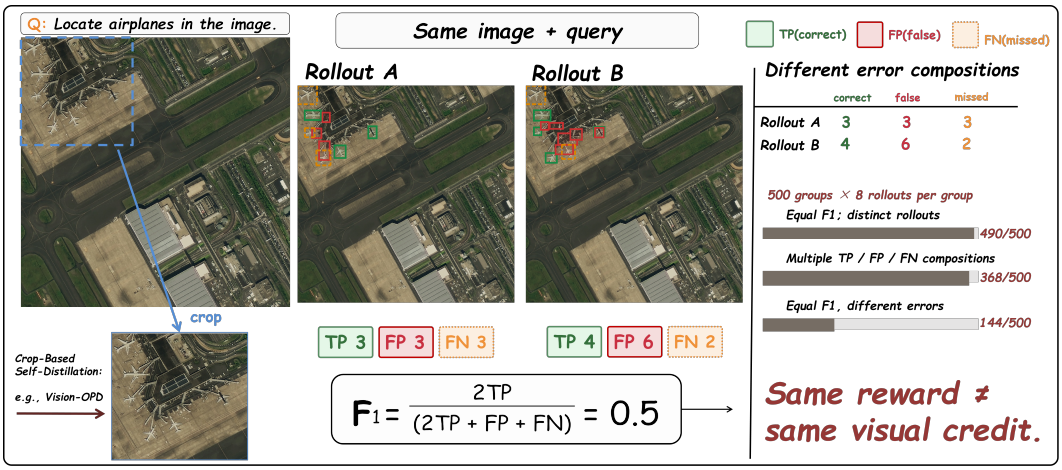

Figure 1: **Same reward, different regional errors.** Crop-based self-distillation (left; Vision-OPD (Yuan et al., 2026b)) uses localized views for supervision, whereas full-image rollouts A and B attain the same F1 with different errors (center). Green, red, and dashed orange boxes mark correct predictions, incorrect predictions, and annotated missed targets, respectively. Group-level statistics (right) motivate supervision beyond a scalar reward.

an additional problem: their coordinates are absent from the response, so adjusting its existing token weights provides no explicit target for recovering them. Finer supervision is therefore needed both to correct what the policy selected and to teach what it failed to predict.

Our insight is that **the regional structure of the policy's errors can supply this missing feedback**. While reward redistribution (Arjona-Medina et al., 2019; Harutyunyan et al., 2019) and process supervision (Lightman et al., 2024; Wang et al., 2024) seek finer signals for intermediate actions and reasoning steps, grounding already exposes useful structure through matching predictions to annotated regions. The verifier identifies correct selections, distractors, and omissions before aggregating them into a reward. Retaining this decomposition makes corrective supervision specific to the mistakes the current policy actually makes. It also makes the source of that supervision inspectable in terms of concrete regional predictions and omissions.

On-policy distillation provides a useful basis for this design by evaluating a sampled response under a more informative condition. Vision-OPD (Yuan et al., 2026b) transfers supervision from localized views to a full-image policy, while RLCSD (Pan et al., 2026) and VCSD (Liang et al., 2026) compare token distributions under paired conditions. We introduce MoVE (**M**odel's **o**wn **V**isual **E**rrors),, which uses verified regional errors to determine both what to contrast and which missing targets to teach. The corrected condition retains matched predictions and restores annotated omissions, while negative conditions preserve errors from the same rollout group. Token-level evidence contrast compares the same response and prefix under these conditions to refine update strength while preserving the outcome advantage's direction. Corrected-region self-distillation supplies the complementary supervision for targets absent from the sampled response.

This design extends existing post-training objectives with **supervision for both erroneous selections and missing targets**. It can be integrated with GRPO and DAPO objectives and combined with on-policy distillation, as instantiated by our extension for fine-grained VQA. The regional conditions are used during training, while inference continues to use the original image and query.

Our contributions are summarized as follows:

- We introduce MoVE, a framework for multimodal post-training that turns verified on-policy regional errors into **corrective supervision beyond a single outcome reward** (Section 3.2).
- We develop **complementary supervision for incorrect selections and omissions**, combining token-level evidence contrast to refine sampled-token credit with corrected-region self-distillation to teach missing targets (Sections 3.3, 3.4).

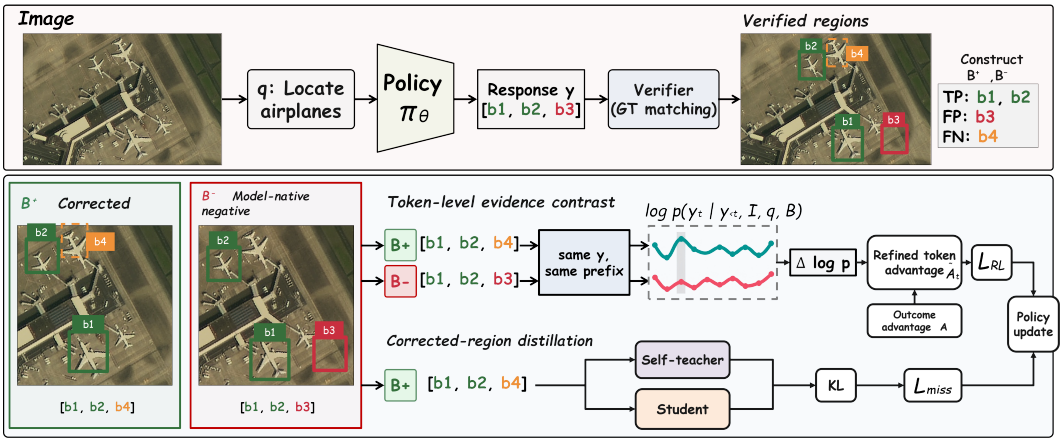

Figure 2: **Overview of MoVE.** Region verification constructs corrected conditions $\mathcal{B}^+$ by retaining matched predictions and restoring missed targets, paired with negative conditions $\mathcal{B}^-$ that preserve rollout errors. Token-level contrast scores the same response under both conditions to refine its credit; corrected-region self-distillation supervises the completed sequence, including omitted coordinates.

- We demonstrate **average grounding gains with both GRPO and DAPO**, and validate corrected-region training across model families and in fine-grained VQA with(Section 4).

2 Related Work

2.1 Fine-grained Visual Understanding and Grounding

Visual grounding connects a language query to localized image content. Fine-grained visual question answering poses a related problem when the answer depends on small or densely distributed evidence. V* (Wu & Xie, 2024) searches high-resolution images for query-relevant regions, while DeepEyes (Zheng et al., 2026) learns to call visual tools such as crop and zoom during reasoning. Both acquire local evidence during inference and therefore require additional visual operations, adding to the resource demands of deploying large models at the network edge (Shao & Li, 2025; An et al., 2026). Region-to-image distillation (Wei et al., 2026) and Vision-OPD (Yuan et al., 2026b) transfer information from localized views into a model that answers from the full image. MoVE learns from regions already generated by the current policy, using their matched, distracting, and missing structure as supervision. Inference continues to use the original image and query.

2.2 Multimodal Reinforcement Post-Training

RL with verifiable rewards is used for multimodal reinforcement post-training. GRPO (Shao et al., 2024) estimates a relative advantage from several rollouts for the same input, and DAPO (Yu et al., 2025) modifies clipping and sampling for more stable optimization. Visual-RFT (Liu et al., 2025), VLM-R1 (Shen et al., 2025), and Perception-R1 (Xiao et al., 2026) apply verifiable rewards to visual question answering, grounding, and multimodal reasoning. PAPO further adds perception-aware regularization to multimodal policy optimization (Wang et al., 2026). Response-level rewards summarize success, while a group-relative outcome advantage does not explicitly distinguish valid and invalid regions within a response.

2.3 On-policy Distillation with Privileged Visual Evidence

On-policy distillation evaluates student-generated prefixes with a teacher distribution, so token-level supervision is applied to states visited by the current policy (Agarwal et al., 2024). In multimodal self-distillation, the teacher can be the same model under a more informative visual condition. Vision-OPD transfers crop-conditioned perception to a full-image policy (Yuan et al., 2026b). V-Zero (Sun et al., 2026), Visual-Advantage OPD (Liu et al., 2026), RP-OPSD (Zhu et al., 2026), and OPD-V (Bi

Algorithm 1 MoVE: regional correction for reinforcement post-training.

Require: Training set $\mathcal{D}$; policy π_θ; rollout group size G; negative-condition budget $K_{\max}$.
for step $= 1, \ldots, M$ **do**
  Sample a batch $\mathcal{S} \subset \mathcal{D}$.
  for all $(I, q, \mathcal{B}^*) \in \mathcal{S}$ **do**
    Sample $\mathcal{G} = \{y^{(i)}\}_{i=1}^{G}$ from $\pi_{\theta_{\text{old}}}$.
    Match predicted and annotated boxes to obtain TP, FP, and FN regions.
    Compute rewards and group-relative advantages $\{A^{(i)}\}_{i=1}^{G}$.
    Construct $\mathcal{B}^+$ and up to $K_{\max}$ negative conditions for selected targets.
    Compute $\widetilde{A}_t$ using Eqs. 2–3; use A when no valid negative exists.
    Form confidence-filtered corrected targets z^+ for $\mathcal{L}_{\text{miss}}$.
  end for
  Update θ with $\mathcal{L}_{\text{MoVE}}$; retain all rollouts in $\mathcal{L}_{\text{RL}}$.
end for

et al., 2026) use contrastive evidence, vision-sensitive token weights, resolution privilege, or paired visual views to decide what should be distilled. RLCSD (Pan et al., 2026) and VCSD (Liang et al., 2026) likewise compare distributions under paired conditions to separate useful supervision from changes introduced by the hint or view itself. For MoVE, the training conditions come from the policy's own region errors, drawing on the broader idea that noise can benefit learning (Li, 2022; Zhang et al., 2025; Yuan et al., 2026a).

3 METHODOLOGY

3.1 OVERVIEW

MOVE uses the region matching behind a grounding reward to construct two complementary learning signals (Figure 2). We first form corrected and error-preserving conditions from on-policy predictions (Section 3.2). **Token-level evidence contrast** uses these conditions to refine credit for the sampled response (Section 3.3), while **corrected-region self-distillation** provides supervision for targets missing from it (Section 3.4). Both objectives update the policy under the original image and query (Section 3.5).

3.2 CONSTRUCTING REGIONAL CONDITIONS FROM ROLLOUTS

We present the formulation in the grounding setting. Given an image I and a query q, the current policy samples a group of responses $\mathcal{G} = \{y^{(i)}\}_{i=1}^{G}$. Each response $y^{(i)} = (y_1^{(i)}, \ldots, y_{|y^{(i)}|}^{(i)})$ yields an ordered box sequence $\widehat{\mathcal{B}}^{(i)}$. The grounding verifier performs one-to-one IoU matching between these predictions and the annotated boxes $\mathcal{B}^*$, identifying matched predictions $\mathcal{B}_{\text{TP}}^{(i)}$, unmatched predictions $\mathcal{B}_{\text{FP}}^{(i)}$, and missed targets $\mathcal{B}_{\text{FN}}^{(i)}$. The matching results determine the outcome reward $r^{(i)}$ and its group-relative advantage $A^{(i)}$. MoVE retains this decomposition to construct regional conditions. Matched and unmatched predictions keep the policy's coordinates and generation order; missed targets come from the annotations.

For a selected response, we omit the rollout index for clarity. The positive condition removes false positives and appends the missed targets to the matched predictions:

$$\mathcal{B}^+ = \mathcal{B}_{\text{TP}} \parallel \mathcal{B}_{\text{FN}}, \qquad \mathcal{N} = \{\mathcal{B}_k^-\}_{k=1}^{K}. \tag{1}$$

Here, $\parallel$ denotes sequence concatenation, and $\mathcal{N}$ contains the available negative conditions. These negatives preserve errors from the same rollout group, including false-positive boxes, incomplete region sequences, and lower-reward box sequences from other responses. In Figure 2, $\mathcal{B}^+$ replaces the incorrect b_3 with the missed b_4, while the illustrated $\mathcal{B}^-$ retains b_3 alongside the matched boxes. The negative condition thus represents a partially incorrect prediction, not necessarily a sequence containing only false positives. Each sequence is serialized using the same regional-condition template. This construction provides a corrected reference and model-generated alternatives for the token-level comparison below.

Table 1: **Main grounding results.** Multi-object grounding is evaluated by Micro-F1 (%) on five benchmarks. Referring-expression grounding is evaluated by Acc0.5 (%) on the RefCOCO series. Light blue marks our method. The best and second-best results are marked within the GRPO and DAPO groups separately.

	Multi-object Grounding (Micro-F1, %)						Referring-expression Grounding (Acc0.5, %)			
Method	**UW**	**Air**	**COCO**	**Obj365**	**GQA**	**Avg.**	**RefCOCO**	**RefCOCO+**	**RefCOCOg**	**Avg.**
Qwen3-VL-2B										
Base	19.11	15.61	48.60	32.01	46.65	32.40	89.30	81.97	85.70	85.66
SFT	56.44	76.52	57.57	49.97	49.47	57.99	88.66	80.90	84.80	84.79
GSPO	59.71	72.03	64.67	55.02	48.23	59.93	89.31	82.05	85.29	85.55
OPSD	23.86	20.18	56.46	40.32	42.12	36.59	89.30	81.74	85.19	85.41
RLSD	59.75	69.46	65.05	53.49	49.74	59.50	89.68	82.06	85.72	85.82
RLCSD	58.77	69.74	64.74	55.06	47.48	59.16	89.26	81.82	85.23	85.44
PGPO	53.58	72.04	63.71	52.60	47.02	57.79	88.90	81.86	85.46	85.41
GRPO	56.79	71.43	63.58	55.46	49.36	59.32	**89.43**	**82.14**	**85.52**	**85.70**
PAPO-GRPO	59.60	69.61	**64.66**	54.07	**52.28**	60.05	88.61	80.99	84.54	84.71
MoVE-GRPO	**60.08**	**73.03**	64.09	**55.70**	49.62	**60.50**	89.35	81.82	85.50	85.55
DAPO	58.92	72.37	63.61	54.28	52.07	60.25	89.23	**82.16**	85.72	85.70
PAPO-DAPO	57.51	70.13	63.52	53.19	50.14	58.90	88.61	80.94	84.93	84.83
MoVE-DAPO	**59.36**	**72.98**	**65.98**	**55.62**	**52.93**	**61.37**	**89.27**	82.14	**85.76**	**85.73**
Qwen3-VL-4B										
Base	36.70	27.55	50.49	48.07	54.63	43.49	87.70	83.58	85.66	85.65
SFT	53.89	75.30	60.79	53.94	55.40	59.86	87.45	83.60	85.60	85.55
GRPO	60.94	73.47	62.59	57.92	**53.66**	61.71	84.15	82.63	83.06	83.28
PAPO-GRPO	61.02	74.24	65.78	56.87	53.59	62.30	83.48	80.48	81.35	81.77
MoVE-GRPO	**61.64**	**74.86**	**65.87**	**58.62**	53.15	**62.83**	**86.78**	**83.33**	**85.01**	**85.04**
DAPO	61.04	72.58	62.93	58.46	**53.05**	61.61	**87.44**	**83.65**	**85.52**	**85.54**
PAPO-DAPO	60.34	72.41	**65.84**	57.56	52.89	61.81	84.73	81.86	82.97	83.19
MoVE-DAPO	**62.66**	**73.90**	64.41	**58.58**	52.78	**62.47**	85.83	82.53	84.11	84.16

3.3 Token-Level Evidence Contrast

The regional conditions identify alternative localizations; we now use them to determine how the sampled response is supported at each token. MoVE evaluates **the same response and prefix** under $\mathcal{B}^+$ and each $\mathcal{B}_k^-$, with the image, query, and condition template shared. Let $\bar{\theta} = \mathrm{stopgrad}(\theta)$ denote the detached current policy and define $p_t(\mathcal{B}) = \pi_{\bar{\theta}}(y_t \mid I, q, \mathcal{B}, y_{<t})$. The token-level evidence contrast is

$$e_t = \log p_t(\mathcal{B}^+) - \log\Big[\tfrac{1}{K}\sum_{k=1}^{K} p_t(\mathcal{B}_k^-)\Big]. \tag{2}$$

A positive e_t means the corrected condition supports y_t more strongly than the negative mixture; a negative value means the opposite. We convert this contrast into a bounded residual and refine the outcome advantage:

$$\delta_t = \alpha \tanh(e_t/\tau), \qquad \widetilde{A}_t = \mathrm{sgn}(A)\max\{0, |A| + \mathrm{sgn}(A)\delta_t\}. \tag{3}$$

Here, α controls the residual magnitude, τ its sensitivity to the contrast, and $\mathrm{sgn}(0) = 0$. The outcome advantage sets the update direction, and the residual adjusts its strength. For $e_t > 0$, the token receives stronger reinforcement when $A > 0$, or weaker penalization when $A < 0$; $e_t < 0$ has the reverse effect, subject to clipping at zero. If no valid negative is available, we use $\widetilde{A}_t = A$. The contrast is evaluated along the on-policy response, as in on-policy self-distillation, but supplies a credit correction to the RL objective rather than a separate distribution-matching loss on that response.

3.4 Corrected-Region Self-Distillation

Evidence contrast only reweights tokens already present in the rollout. **A missed target has no coordinate tokens in** y, so this update supplies no explicit target for the omitted region. We therefore

serialize $\mathcal{B}^+$ into a corrected sequence z^+ and distill it into the policy under the original full-image input. Let $c^+ = c(\mathcal{B}^+)$ denote the corresponding training-time condition. The detached policy provides $P_l^+ = \pi_{\bar{\theta}}(\cdot \mid I, q, c^+, z^+_{<l})$, and the full-image student provides $P_l^{\text{full}} = \pi_\theta(\cdot \mid I, q, z^+_{<l})$. Their forward KL defines

$$\mathcal{L}_{\text{miss}} = \tfrac{1}{|z^+|} \sum_{l=1}^{|z^+|} \mathrm{KL}\big(P_l^+ \,\|\, P_l^{\text{full}}\big)\,. \tag{4}$$

This branch uses the corrected sequence rather than replaying y, extending supervision to coordinates absent from the on-policy response. Direct cross-entropy on z^+ would supervise only the selected target token; the KL objective transfers the teacher distribution over the vocabulary at each corrected prefix. We apply this loss to confidence-filtered corrected targets.

3.5 Training Objective

We combine credit refinement for sampled tokens with distillation of the corrected sequence. For each sampled token, define the importance ratio

$$\rho_t(\theta) = \tfrac{\pi_\theta(y_t|I,q,y_{<t})}{\pi_{\theta_{\text{old}}}(y_t|I,q,y_{<t})}. \tag{5}$$

Taking GRPO as an example, the clipped policy objective is

$$\mathcal{L}_{\text{RL}} = -\mathbb{E}_{y\sim\pi_{\theta_{\text{old}}}}\left[\tfrac{1}{|y|}\sum_{t=1}^{|y|}\min\Big(\rho_t(\theta)\widetilde{A}_t,\ \mathrm{clip}(\rho_t(\theta),1-\epsilon,1+\epsilon)\widetilde{A}_t\Big)\right]. \tag{6}$$

The full objective is

$$\mathcal{L}_{\text{MoVE}} = \mathcal{L}_{\text{RL}} + \lambda_{\text{miss}}\mathcal{L}_{\text{miss}}, \tag{7}$$

where λ_{miss} balances the two learning signals. The policy continues to generate from the original image and query; the regional conditions are only used to compute training targets. Algorithm 1 summarizes the grounding procedure.

Extension to region-sensitive VQA. MoVE can also use region predictions to refine supervision for a textual answer. During training, an auxiliary grounding output makes the regions supporting the answer explicit. Where suitable regional supervision is available, these coordinates can define corrected and erroneous conditions for scoring the same answer tokens. The resulting contrast refines answer-token credit, while a regional teacher transfers information from the informative view to the full-image policy. The form of this distillation target depends on the task: the VQA instantiation uses answer distributions rather than a corrected box sequence. Section 4.3 describes the implementation.

4 Experiments

We organize the experiments around the following four questions. **Q1:** Does MoVE improve visual grounding over SFT and outcome-RL baselines? **Q2:** Can the same regional correction improve fine-grained visual question answering? **Q3:** What is gained by restoring missed regions, and what is added by contrasting them with the model's own errors? **Q4:** Does MoVE generalize across model families and scales?

4.1 Experimental Setup

Models and training data. Our main grounding experiments use Qwen3-VL-2B-Instruct and Qwen3-VL-4B-Instruct (Bai et al., 2025). MSVQA covers visual question answering and grounding across four scenarios with distinct backgrounds and viewpoints, underwater, high-altitude, low-altitude, and indoor imagery (Jiang et al., 2025). We use its underwater and high-altitude subsets, whose changes in scale, viewpoint, and appearance provide two complementary multi-object settings. The training set contains 6,000 samples, 2,000 examples from each MSVQA subset and 2,000 examples from COCO2017 (Lin et al., 2014). We construct a validation set from the same three sources.

Table 2: **Fine-grained VQA results under the Vision-OPD protocol.** Accuracy (%) is reported for Qwen3.5-4B and Qwen3.5-9B. Light blue marks our method, and bold marks the best result.

Method	Fine-grained visual tasks				Holdout tasks			
	V* Bench	ZoomBench	HR-Bench 4K	HR-Bench 8K	MMVP	CV-Bench	MMStar	POPE
Qwen3.5-4B								
Vanilla	84.29	47.69	84.38	80.13	76.67	87.13	78.53	88.28
SFT on Self-Teacher	78.01	54.67	79.75	76.38	78.33	85.70	68.40	87.48
GRPO	83.77	55.38	82.63	78.25	79.33	87.24	70.60	86.37
DAPO	84.82	55.74	84.00	78.00	79.33	86.95	72.27	86.62
OPSD	85.34	53.85	82.25	79.38	78.68	87.27	75.07	88.83
Vision-OPD	**92.15**	59.76	84.50	80.38	**79.67**	87.27	**79.60**	**89.14**
MoVE	90.58	**61.30**	**86.38**	**81.75**	78.33	**88.05**	78.47	88.24
Qwen3.5-9B								
Vanilla	82.72	52.07	85.75	80.63	83.33	88.29	83.07	88.88
SFT on Self-Teacher	82.20	58.46	83.50	80.25	80.67	87.97	73.33	87.47
GRPO	85.34	57.51	86.88	83.25	81.67	87.78	73.40	87.78
DAPO	88.48	55.62	86.25	84.13	79.33	87.30	75.67	87.54
OPSD	89.53	57.51	84.00	81.25	80.67	87.45	79.13	87.48
Vision-OPD	**94.76**	**65.80**	88.13	85.50	**83.67**	88.40	**83.20**	**89.13**
MoVE	92.15	60.83	**90.25**	**86.50**	79.67	**89.88**	78.20	85.27

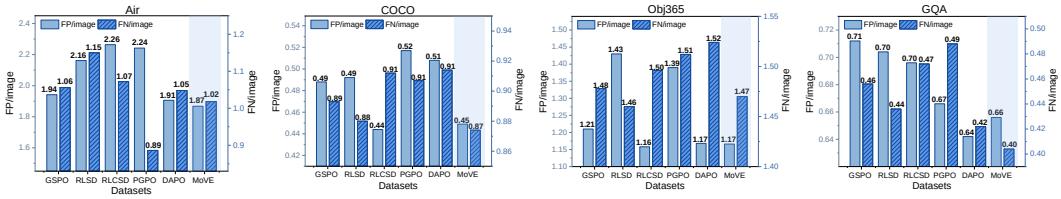

Figure 3: **False positives and missed targets across grounding methods.** Solid and hatched bars denote average false positives and missed targets, respectively. The highlighted MoVE bars remain near the low-error end on both measures.

Evaluation. The held-out MSVQA and COCO2017 splits measure in-domain multi-object grounding. Objects365 (Shao et al., 2019) and GQA (Hudson & Manning, 2019) test transfer to new image and query distributions. RefCOCO (Yu et al., 2016), RefCOCO+ (Yu et al., 2016), and RefCOCOg (Mao et al., 2016) provide a separate referring-expression evaluation. We report Micro-F1 for multi-object grounding and Acc@0.5 for the RefCOCO series. Fine-grained VQA follows the benchmark metrics and evaluation protocol used by Vision-OPD (Yuan et al., 2026b).

Comparisons. We compare Base and SFT with outcome-RL and distillation baselines, including GRPO (Shao et al., 2024), GSPO (Zheng et al., 2025), DAPO (Yu et al., 2025), OPSD (Zhao et al., 2026), RLSD (Yang et al., 2026), RLCSD (Pan et al., 2026), and PGPO (Ye et al., 2026). PAPO (Wang et al., 2026) and MoVE are evaluated with both GRPO and DAPO clipping, using the same model, training data, and evaluation setting within each comparison.

4.2 Main Grounding Results

Table 1 compares MoVE with the policy objective it extends. Under both GRPO and DAPO, **MoVE improves the average multi-object grounding result for the 2B and 4B models**. The gains cover the underwater, aerial, and COCO splits, showing improved localization within the three training domains. MoVE also improves Objects365 in all four matched comparisons, extending the benefit to a dataset not used for training.

The additional RL objectives lead to gains on different subsets, but these improvements are not consistent across evaluation settings. In contrast, MoVE achieves a positive average gain in both the matched GRPO and DAPO comparisons. The improvement is most consistent on multi-object grounding, where a single response may contain multiple region predictions with different levels of correctness. On the RefCOCO series, which evaluates single-object referring expressions, the

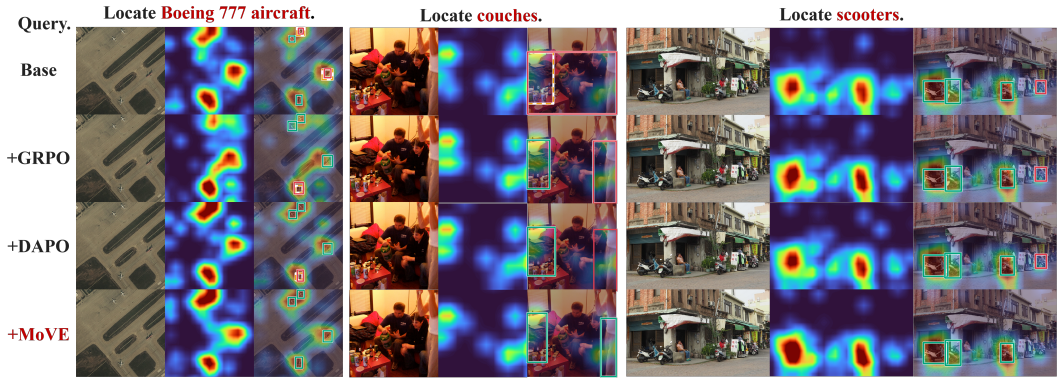

Figure 4: **Qualitative comparison of regional predictions and attention maps.** Rows show Base, GRPO, DAPO, and full MoVE, from top to bottom. Each example group contains the input image, an attention map, and their overlay with regional boxes. Green, red, and dashed orange boxes indicate matched predictions, false positives, and missed targets.

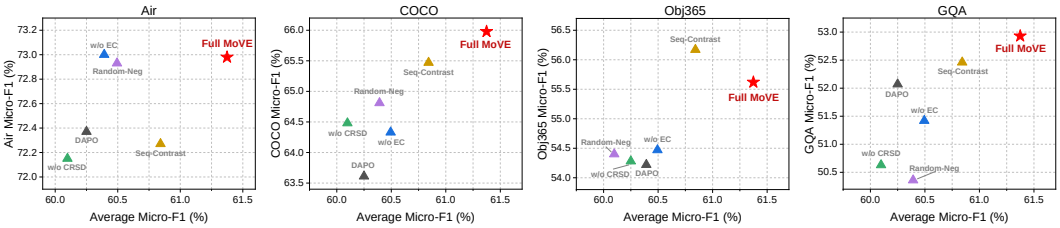

Figure 5: **Contributions of regional correction, negative conditions, and token-level contrast.** The panels show Air, COCO, Obj365, and GQA from left to right. Each point reports the five-benchmark average on the horizontal axis and the Micro-F1 on the benchmark shown in its panel on the vertical axis. The complete model is highlighted in red.

methods perform more similarly and the gains are less consistent. This contrast is consistent with MoVE's focus on correcting heterogeneous regional evidence within a response. The corresponding error profile is shown in Figure 3; across these splits, MoVE stays near the low-error end for both false positives and missed targets. Figure 4 provides a qualitative comparison of regional predictions and spatial maps for the same queries.

4.3 Fine-grained Visual Question Answering

We follow the training and evaluation protocol of Vision-OPD and augment each answer rollout with the regions that support its prediction. The predicted coordinates are used to construct a corrected regional view and an erroneous view rendered from a region proposed by the frozen base policy. An EMA teacher evaluates the same answer tokens under both views. The corrected view provides the Vision-OPD answer-distillation target, while the positive–negative log-likelihood difference refines the group-relative advantage according to Eq. 3. This instantiation distills answer distributions from regional views. Regional inputs are removed after training, and all benchmarks are evaluated from the full image and the original question.

Table 2 places MoVE among the top two methods on all four fine-grained visual benchmarks at both model scales. It achieves the best results on both HR-Bench resolutions and the held-out CV-Bench at 4B and 9B, and also leads ZoomBench at 4B. These results show that rollout-derived regional correction can support textual answer prediction beyond explicit grounding, with consistent gains on fine-grained task.

Table 3: **Generalization of region correction across model families and scales.** Multi-object columns report Micro-F1 (%); RefCOCO columns report Acc@0.5 (%). Green ↑ and red ↓ values denote absolute changes from the corresponding base model. Each trained row uses MoVE's corrected-region branch with GRPO. Bold marks the higher score within each base–trained pair, and light blue marks the trained row.

Base model	Multi-object grounding (Micro-F1)						Referring-expression grounding (Acc@0.5)			
	UW	Air	COCO	Obj365	GQA	Avg.	RefCOCO	RefCOCO+	RefCOCOg	Avg.
Qwen3-VL-2B	19.11	15.61	48.60	32.01	46.65	32.40	89.30	**81.97**	**85.70**	**85.66**
+ MoVE	**60.08** ↑40.97	**73.03** ↑57.42	**64.09** ↑15.49	**55.70** ↑23.69	**49.62** ↑2.97	**60.50** ↑28.10	**89.35** ↑0.05	81.82 ↓0.15	85.50 ↓0.20	85.55 ↓0.11
Qwen3-VL-4B	36.70	27.55	50.49	48.07	**54.63**	43.49	**87.70**	**83.58**	**85.66**	**85.65**
+ MoVE	**61.64** ↑24.94	**74.86** ↑47.31	**65.87** ↑15.38	**58.62** ↑10.55	53.15 ↓1.48	**62.83** ↑19.34	86.78 ↓0.92	83.33 ↓0.25	85.01 ↓0.65	85.04 ↓0.61
Qwen3-VL-8B	31.72	29.40	51.14	43.06	56.35	42.33	**92.39**	87.52	89.05	**89.65**
+ MoVE	**61.41** ↑29.69	**74.30** ↑44.90	**62.75** ↑11.61	**54.13** ↑11.07	**56.69** ↑0.34	**61.86** ↑19.53	92.10 ↓0.29	**87.57** ↑0.05	**89.10** ↑0.05	89.59 ↓0.06
InternVL3.5-2B	21.48	23.79	39.57	16.02	31.73	26.52	21.47	15.85	24.24	20.52
+ MoVE	**34.50** ↑13.02	**34.07** ↑10.28	**47.98** ↑8.41	**28.92** ↑12.90	**43.29** ↑11.56	**37.75** ↑11.23	**47.22** ↑25.75	**42.63** ↑26.78	**52.96** ↑28.72	**47.60** ↑27.08
InternVL3.5-4B	42.11	30.39	60.93	39.90	41.60	42.99	66.72	55.37	68.73	63.61
+ MoVE	**47.19** ↑5.09	**33.78** ↑3.39	**63.32** ↑2.39	**43.15** ↑3.25	**43.80** ↑2.20	**46.25** ↑3.26	**77.48** ↑10.76	**65.55** ↑10.18	**76.82** ↑8.09	**73.28** ↑9.68
InternVL3.5-8B	39.13	27.05	58.28	37.64	40.20	40.46	61.65	49.80	57.76	56.40
+ MoVE	**44.79** ↑5.66	**28.14** ↑1.09	**59.56** ↑1.28	**41.02** ↑3.38	**43.80** ↑3.60	**43.46** ↑3.00	**70.00** ↑8.35	**58.37** ↑8.57	**65.30** ↑7.54	**64.56** ↑8.16
Qwen3.5-2B	28.93	37.11	47.68	34.40	41.33	37.89	88.87	82.20	86.11	85.73
+ MoVE	**63.63** ↑34.70	**78.20** ↑41.09	**63.67** ↑15.99	**52.87** ↑18.47	**47.60** ↑6.28	**61.19** ↑23.31	**89.57** ↑0.70	**82.57** ↑0.37	**87.03** ↑0.92	**86.39** ↑0.66
Qwen3.5-4B	48.40	41.22	59.67	51.05	**57.04**	51.48	**89.18**	**83.80**	**85.68**	**86.22**
+ MoVE	**56.24** ↑7.83	**67.42** ↑26.20	**62.31** ↑2.63	**53.14** ↑2.09	56.98 ↓0.06	**59.22** ↑7.74	88.70 ↓0.48	83.61 ↓0.19	85.21 ↓0.47	85.84 ↓0.38
Qwen3.5-9B	47.46	46.02	60.65	**54.57**	**53.64**	52.47	90.11	**85.69**	**87.97**	**87.92**
+ MoVE	**55.61** ↑8.15	**52.80** ↑6.78	**61.43** ↑0.78	54.19 ↓0.38	48.57 ↓5.07	**54.52** ↑2.05	**90.20** ↑0.09	85.48 ↓0.21	87.95 ↓0.02	87.88 ↓0.04

4.4 Ablation Study

Figure 5 examines the contributions of MoVE's learning objectives, negative conditions, and contrast granularity using Qwen3-VL-2B with DAPO. We denote token-level evidence contrast by EC and corrected-region self-distillation by CRSD; Full MoVE combines both objectives with DAPO. **The two learning objectives work best together.** The variants w/o EC and w/o CRSD remove the corresponding objective while retaining the other. Full MoVE achieves the highest average Micro-F1 of 61.37, exceeding these two variants by 0.88 and 1.27 points, respectively. This comparison supports their complementary roles. EC refines supervision for tokens already generated, while CRSD supplies targets absent from the sampled response.

4.5 Generalization Across Model Families and Scales

Table 3 extends the corrected-region branch to nine base models across three model families and multiple parameter scales. MoVE brings a positive gain in the five-set grounding average for every model, showing that the improvement is not tied to a specific backbone or scale. Its advantage is most consistent on multi-object grounding, where regional correction directly addresses mixed-quality predictions within the same response.

5 Conclusion

We present MoVE, a framework that extends multimodal reinforcement post-training with **explicit supervision for regional errors and omissions**. By learning from the policy's own verified predictions, MoVE connects credit refinement within a response with supervision for targets the response did not produce. Matched grounding experiments show average gains with both GRPO and DAPO at two model scales, and the Vision-OPD extension demonstrates that regional correction also benefits fine-grained VQA. Together, these results establish regional correction as a complementary source of supervision for existing post-training objectives, addressing both the visual decisions a policy gets wrong and the targets it overlooks.

AI USE STATEMENT

Generative AI tools were used to provide feedback on the methodological framing and experimental design and translate and polish portions of the manuscript. They were also used to improve the manuscript's readability and organization and modify code for the research pipeline and experimental visualizations. The authors reviewed the AI-assisted material retained in the submission. AI-suggested references were checked against their original sources, code changes were inspected and evaluated in the relevant pipelines, and reported results were checked against experimental records. The authors take responsibility for the final content of this work, including its text, claims, analyses, figures, and software artifacts.

REPRODUCIBILITY STATEMENT

Reproducibility details are provided in the main text and appendices. Section 3 presents MoVE and its optimization objectives, while Section 4.1 summarizes the models, datasets, metrics, and comparison protocols. Appendices A–D describe the datasets and evaluation, implementation, regional-condition construction, and baseline reproduction protocols. Appendices E–G provide additional analyses, cross-model results, and computational costs. We plan to release the training and evaluation code, along with data-processing scripts and data splits.

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.

Hongjun An, Wenhan Hu, Sida Huang, Siqi Huang, Ruanjun Li, Yuanzhi Liang, Jiawei Shao, Yiliang Song, Zihan Wang, Cheng Yuan, et al. Ai flow: Perspectives, scenarios, and approaches. *Vicinagearth*, 3(1):1, 2026.

Jose A Arjona-Medina, Michael Gillhofer, Michael Widrich, Thomas Unterthiner, Johannes Brandstetter, and Sepp Hochreiter. Rudder: Return decomposition for delayed rewards. volume 32, 2019.

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, 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.

Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, et al. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling, 2024.

Matt Deitke, Christopher Clark, Sangho Lee, Rohun Tripathi, Yue Yang, Jae Sung Park, Mohammadreza Salehi, Niklas Muennighoff, Kyle Lo, Luca Soldaini, et al. Molmo and pixmo: Open weights and open data for state-of-the-art vision-language models. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 91–104. IEEE, 2025.

Anna Harutyunyan, Will Dabney, Thomas Mesnard, Mohammad Gheshlaghi Azar, Bilal Piot, Nicolas Heess, Hado P van Hasselt, Gregory Wayne, Satinder Singh, Doina Precup, et al. Hindsight credit assignment. volume 32, 2019.

Drew A Hudson and Christopher D Manning. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6693–6702. IEEE, 2019.

Kai Jiang, Siqi Huang, Xiangyu Chen, Jiawei Shao, Hongyuan Zhang, Ping Luo, and Xuelong Li. Multimodal continual learning with mllms from multi-scenario perspectives. *arXiv preprint arXiv:2511.18507*, 2025.

Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, et al. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*, 2024.

Xuelong Li. Positive-incentive noise. *IEEE Transactions on Neural Networks and Learning Systems*, 35(6):8708–8714, 2022.

Yijun Liang, Yunjie Tian, Yijiang Li, Yuqi Jia, Furong Huang, Tianyi Zhou, and Di Fu. Visual contrastive self-distillation. 2026. URL https://arxiv.org/abs/2607.21556.

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.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pp. 740–755. Springer, 2014.

Ruiqi Liu, Xiaolei Lv, Gengsheng Li, Ximo Zhu, Zhiheng Wang, Zhengbo Zhang, Junkai Chen, Zhiheng Li, Bo Li, Jun Gao, et al. Visual-advantage on-policy distillation for vision-language models. *arXiv preprint arXiv:2605.21924*, 2026.

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

Junhua Mao, Jonathan Huang, Alexander Toshev, Oana Camburu, Alan L Yuille, and Kevin Murphy. Generation and comprehension of unambiguous object descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 11–20, 2016.

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.

Jiawei Shao and Xuelong Li. Ai flow at the network edge. *IEEE Network*, 40(1):330–336, 2025.

Shuai Shao, Zeming Li, Tianyuan Zhang, Chao Peng, Gang Yu, Xiangyu Zhang, Jing Li, and Jian Sun. Objects365: A large-scale, high-quality dataset for object detection. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 8429–8438. IEEE, 2019.

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

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

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.

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 (Volume 1: Long Papers)*, pp. 9426–9439, 2024.

Zhenhailong Wang, Xuehang Guo, Sofia Stoica, Haiyang Xu, Hongru Wang, Hyeonjeong Ha, Xiusi Chen, Yangyi Chen, Ming Yan, Fei Huang, et al. Perception-aware policy optimization for multimodal reasoning. In *International Conference on Learning Representations*, volume 2026, pp. 80502–80530, 2026.

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.

Tong Xiao, Xin Xu, Zhenya Huang, Hongyu Gao, Quan Liu, Qi Liu, and Enhong Chen. Perception-r1: Advancing multimodal reasoning capabilities of mllms via visual perception reward. In *International Conference on Learning Representations*, volume 2026, pp. 26868–26898, 2026.

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.

Zekai Ye, Qiming Li, Xiaocheng Feng, Ruihan Chen, Ziming Li, Haoyu Ren, Kun Chen, Dandan Tu, and Bing Qin. Not all tokens see equally: Perception-grounded policy optimization for large vision-language models. *arXiv preprint arXiv:2604.01840*, 2026.

Licheng Yu, Patrick Poirson, Shan Yang, Alexander C Berg, and Tamara L Berg. Modeling context in referring expressions. In *European conference on computer vision*, pp. 69–85. Springer, 2016.

Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Weinan Dai, Tiantian Fan, Gaohong Liu, Lingjun Liu, Xin Liu, Haibin Lin, Zhiqi Lin, Bole Ma, Guangming Sheng, Yuxuan Tong, Chi Zhang, Mofan Zhang, Wang Zhang, Hang Zhu, Jinhua Zhu, Jiaze Chen, Jiangjie Chen, Chengyi Wang, Hongli Yu, Yuxuan Song, Xiangpeng Wei, Hao Zhou, Jingjing Liu, Wei-Ying Ma, Ya-Qin Zhang, Lin Yan, Mu Qiao, Yonghui Wu, and Mingxuan Wang. Dapo: An open-source llm reinforcement learning system at scale, 2025. URL https://arxiv.org/abs/2503.14476.

Cheng Yuan, Zhenyu Jia, Jiawei Shao, and Xuelong Li. Enhancing neural video compression of static scenes with positive-incentive noise. *arXiv preprint arXiv:2603.06095*, 2026a.

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*, 2026b.

Hongyuan Zhang, Sida Huang, Yubin Guo, and Xuelong Li. Variational positive-incentive noise: How noise benefits models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47 (9):8313–8320, 2025.

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.

Chujie Zheng, Shixuan Liu, Mingze Li, Xiong-Hui Chen, Bowen Yu, Chang Gao, Kai Dang, Yuqiong Liu, Rui Men, An Yang, Jingren Zhou, and Junyang Lin. Group sequence policy optimization. 2025. URL https://arxiv.org/abs/2507.18071.

Ziwei Zheng, Minghao Yang, Jack Hong, Chenxiao Zhao, Guohai Xu, Le Yang, and Chao Shen. Deepeyes: Incentivizing" thinking with images" via reinforcement learning. 2026:126775–126798, 2026.

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.

APPENDIX

A DATASET AND EVALUATION DETAILS

We evaluate multi-object grounding, referring-expression grounding, and region-sensitive VQA. For multi-object grounding, each query specifies the target category and the parser extracts every coordinate tuple from the response. Valid predicted and annotated boxes are matched one to one in descending IoU order at threshold 0.5. Every valid prediction left unmatched is counted as a false positive, and every unmatched annotation is counted as a false negative. We pool these counts over all examples within each dataset before computing Micro-F1:

$$F_1 = \frac{2\mathrm{TP}}{2\mathrm{TP} + \mathrm{FP} + \mathrm{FN}}. \tag{A1}$$

The five-dataset average is the arithmetic mean of the five Micro-F1 scores, not the F1 obtained by pooling different datasets. FP/image and FN/image measure the number of unmatched predictions and targets per evaluated example.

RefCOCO, RefCOCO+, and RefCOCOg are evaluated on their validation splits with the official referring-expression protocol and Acc@0.5. The VQA numbers follow the benchmark protocol used by Vision-OPD; the model receives the full image and the original question at test time. Region coordinates used to build training conditions are removed before evaluation.

Table A1: **Dataset and evaluation details.**

Dataset	Evaluation setting	Train	Eval	Metric
MSVQA–Underwater	In-domain multi-object	2,000	500	Micro-F1
MSVQA–High-altitude	In-domain multi-object	2,000	500	Micro-F1
COCO2017	In-domain multi-object	2,000	1,000	Micro-F1
Objects365	Cross-dataset multi-object	–	500	Micro-F1
GQA	Cross-dataset multi-object	–	500	Micro-F1
RefCOCO	Referring expression, val	–	10,834	Acc@0.5
RefCOCO+	Referring expression, val	–	10,758	Acc@0.5
RefCOCOg	Referring expression, val	–	4,896	Acc@0.5
Vision-OPD	Region-sensitive VQA, official test	–	Official test	Official metrics

B IMPLEMENTATION DETAILS

B.1 GROUNDING SETUP

MoVE retains the MLLM backbone and its visual-token interface. Grounding runs use four NVIDIA H100 80GB GPUs with DeepSpeed ZeRO-2. For Qwen3-VL-2B and 4B, we load the image processor from the corresponding pretrained checkpoint. Both processors use 16×16 image patches and 2×2 spatial merging, so resized image dimensions are aligned to multiples of 32. The processor configuration specifies pixel budgets of 65,536 and 16,777,216 for the minimum and maximum image area. The visual-token count depends on the resized image dimensions, with one token per merged 32×32 spatial unit.

Within each backbone, the comparisons share the training split, image preprocessing, output parser, and evaluation files. Outcome-RL runs also share the grounding reward and rollout configuration. The grounding experiments use Qwen3-VL-2B-Instruct and Qwen3-VL-4B-Instruct. We train for two epochs with group size $G = 8$ and bf16 mixed precision. Coordinates are normalized to $[0, 1000]$. The verifier uses one-to-one IoU matching with threshold 0.5.

For GRPO-style runs, the policy ratio is clipped to $[0.80, 1.20]$. DAPO runs use the asymmetric interval $[0.80, 1.28]$ from the matched baseline implementation. MoVE uses $\alpha = 0.25$ for the contrastive residual, temperature $\tau = 0.05$, and $\lambda_{\mathrm{miss}} = 10^{-3}$ for corrected-region distillation. The negative candidate budget is $K = 4$.

B.2 FINE-GRAINED VQA SETUP

The VQA extension follows its existing training protocol. The EMA teacher uses update rate 0.05, each input samples $G = 8$ answer rollouts, the regional teacher retains TopK $= 100$ visual units, and training runs 1 epoch. MoVE's residual scale is 0.20 with temperature 0.50. The auxiliary region output is used only to construct training views; it is not part of the inference interface.

The regional coordinates provide training views for answer supervision. We retain Vision-OPD's positive view and construct the negative view from a region proposed by the frozen base policy. The EMA teacher scores the same answer under both views. The positive view supplies the distillation target; their likelihood contrast adjusts the answer-token advantages. This experiment uses regional views rather than the corrected box-sequence target used in grounding. Test inputs remain the full image and the original question.

C REGIONAL CONDITION CONSTRUCTION

We construct evidence conditions from partially correct rollouts, using matched predictions as the basis for regional correction. For evidence contrast, we select up to two parseable responses per rollout group with $0.4 \leq \mathrm{F1} < 1$, at least one matched prediction, and $\mathrm{FP} + \mathrm{FN} \leq 4$. Corrected-region distillation uses one target per optimization microbatch, preferably from the same candidates. If none is available, we relax the error-count constraint and then select the highest-F1 parseable response. All sampled rollouts still contribute to the outcome-RL objective.

Negative conditions preserve errors made by the policy. False-positive variants reintroduce up to four incorrect boxes in their generation order; conditions drawn from other rollouts in the group retain the ordered box sequences and are ranked by rollout statistics. After deduplication, we keep at most $K = 4$ negative conditions per target. Responses without a valid negative retain their original outcome advantage. For corrected-region distillation, the teacher NLL threshold $\gamma = 1.2$ selects the corrected sequences used for supervision; the distillation loss is omitted when none passes this threshold.

D BASELINE REPRODUCTION PROTOCOL

The grounding baselines use the shared training split described in Appendix A and the same evaluation pipeline. Comparisons are made within each backbone. GSPO, OPSD, RLSD, RLCSD, and PGPO are evaluated on Qwen3-VL-2B and Qwen3-VL-4B. The following describes how each baseline is instantiated for grounding.

Base and SFT. Base evaluates the pretrained instruction model directly. SFT trains the same model to predict the annotated target response with token-level cross-entropy.

GRPO, DAPO, and GSPO. GRPO uses group-relative outcome advantages with token-level importance ratios and symmetric clipping. DAPO uses the asymmetric clipping interval specified above. GSPO instead computes a length-normalized sequence importance ratio and clips it at the response level. These runs share the grounding reward and rollout group size.

PAPO. PAPO adds a perception regularizer based on the response-likelihood difference between the original and perturbed visual inputs. We combine this regularizer with GRPO and DAPO to obtain the two PAPO baselines in the main table.

OPSD. The student generates responses from the original image and query. A frozen copy of the initial model receives the annotated target response as an additional hint and supplies token distributions along the student-generated prefixes. This run optimizes distillation without an outcome-policy loss.

RLSD. RLSD evaluates each sampled response with the current policy conditioned on the annotated target response. The teacher–student log-likelihood difference reweights the group-relative advantage

at each token. The weights are clipped to $[0.8, 1.2]$ and mixed with the original advantage using coefficient 0.5.

RLCSD. RLCSD constructs positive and negative hints from higher- and lower-scoring responses in the same rollout group. The same reference-response template is used on both sides. Their token-level likelihood contrast supplies a bounded residual to the outcome advantage. Our grounding implementation uses up to four negative responses and retains the sign of the original advantage.

PGPO. PGPO estimates token-level visual dependence by comparing predictive distributions with and without the image. Our implementation computes this divergence over the top 100 vocabulary entries and a residual probability bucket, then uses the normalized dependence scores to reweight the DAPO advantage. The weighting threshold is 0.4 and the boost coefficient is 2.0.

E Training-Seed Sensitivity

We repeat the Qwen3-VL-2B comparison between DAPO and MoVE-DAPO with training seeds 42, 43, and 44, using the same data split and evaluation protocol. Table A2 reports the mean with sample standard deviation across seeds; the seed-42 runs correspond to Table 1. For each run, Avg. is the equal-weight mean of the five dataset scores, and its standard deviation is computed across these per-run averages.

Table A2: **Grounding results across three training seeds.** Scores are Micro-F1 (%); results report mean ± sample standard deviation over three training seeds, with the higher mean in bold.

Method	UW	Air	COCO	Obj365	GQA	Avg.
DAPO	$\mathbf{59.10} \pm 0.17$	72.42 ± 0.31	64.13 ± 0.46	53.90 ± 0.40	$\mathbf{51.57} \pm 1.09$	60.23 ± 0.31
MoVE-DAPO	58.80 ± 0.49	$\mathbf{73.00} \pm 0.21$	$\mathbf{65.51} \pm 0.42$	$\mathbf{54.87} \pm 0.75$	50.71 ± 2.10	$\mathbf{60.58} \pm 0.76$

F Ablation Study

Table A3 brings together the component ablations in Section 4.4 and the comparisons with alternative supervision. All variants use Qwen3-VL-2B with the same training data, two-epoch budget, and DAPO objective.

Learning objectives. EC denotes token-level evidence contrast, and CRSD denotes corrected-region self-distillation, following Figure 5. The w/o EC variant retains DAPO and CRSD, while w/o CRSD retains DAPO with contrast-adjusted token advantages. These variants separate the refinement of sampled-token updates from the supervision supplied by corrected targets.

Negative conditions and contrast granularity. Random-Neg and Seq-Contrast retain both objectives. Random-Neg replaces model-native negative conditions with random negatives; Seq-Contrast applies a single contrastive correction to the entire response.

Cross-entropy on corrected targets. This variant replaces the KL loss in CRSD with cross-entropy on the same target sequence, retaining the teacher-confidence gate, full-image student input, and token normalization. Both losses use coefficient 10^{-3}, so the comparison tests the loss substitution at a fixed coefficient.

Direct TP/FP weighting. This variant replaces likelihood contrast with a residual of $+0.25$ for TP coordinate-number tokens and -0.25 for FP coordinate-number tokens. It retains MoVE's target eligibility, valid-negative requirement, sign-preserving projection, and CRSD; other tokens keep the outcome advantage.

RLCSD with CRSD. This control combines RLCSD's response-level positive and negative hints with the corrected-region KL objective and its confidence gate. It uses DAPO clipping while retaining RLCSD's hint selection and contrast settings.

Table A3: **Component ablations and alternative supervision with Qwen3-VL-2B and DAPO.** Scores are Micro-F1 (%); Avg. is the arithmetic mean across the five datasets. EC denotes token-level evidence contrast, and CRSD denotes corrected-region self-distillation. Bold and underlined values mark the best and second-best results across all ablation variants, including Full MoVE and excluding the training baselines. Light blue marks Full MoVE.

Setting	UW	Air	COCO	Obj365	GQA	Avg.
Training baselines						
Base	19.11	15.61	48.60	32.01	46.65	32.40
GRPO	56.79	71.43	63.58	55.46	49.36	59.32
DAPO	58.92	72.37	63.61	54.28	52.07	60.25
Component ablations						
w/o EC	59.32	72.93	64.33	54.47	51.42	60.49
w/o CRSD	58.83	72.15	64.48	54.40	50.63	60.10
Random-Neg	**59.57**	**73.00**	64.81	54.22	50.36	60.39
Seq-Contrast	57.84	72.27	<u>65.47</u>	**56.17**	<u>52.46</u>	<u>60.84</u>
Alternative supervision						
MoVE with corrected-target CE	55.66	72.38	64.78	55.47	51.53	59.96
Direct TP/FP weighting + CRSD	56.17	72.25	64.20	53.47	50.43	59.30
RLCSD + CRSD	56.83	69.92	64.22	54.53	50.88	59.28
Full MoVE	<u>59.36</u>	<u>72.98</u>	**65.98**	<u>55.62</u>	**52.93**	**61.37**

Full MoVE achieves the best five-dataset average and the highest scores on COCO and GQA. Among the alternative-supervision controls, MoVE attains the highest score on all five datasets. Replacing KL with cross-entropy lowers the average by 1.41 points, favoring distributional supervision at the tested coefficient. Direct region weighting and RLCSD with CRSD reach 59.30 and 59.28, respectively, so neither recovers the full result by retaining corrected targets alone.

G Training Efficiency

We measure training cost on Qwen3-VL-2B under the same setting used by the ablation study: the shared 6,000-example training split, four GPUs, 2 epochs, eight rollouts per prompt, and bfloat16 training. DAPO and all evidence-learning variants use the same asymmetric clipping objective and batch configuration. Wall time includes rollout generation, reward evaluation, and all auxiliary replay and distillation passes. Table A4 reports measurements from one completed end-to-end run for each setting on the same four-GPU worker.

Table A4: **Training efficiency on Qwen3-VL-2B.**

Method	Avg. Micro-F1	Time (h)	Sec./step	GPU-hours	Peak mem. (GiB)
DAPO	60.25	7.56	72.34	30.22	40.44
w/o EC	60.49	7.28	69.70	29.12	40.48
Seq-Contrast	60.84	7.91	75.78	31.66	40.44
Full MoVE	**61.37**	7.73	74.04	30.93	40.33

MoVE improves average Micro-F1 by 1.12 points over DAPO with only 2.3% additional training time. It reuses the region matching performed for reward computation and evaluates existing responses under different evidence conditions through teacher forcing. These updates require no additional autoregressive generation, keeping the training overhead modest.

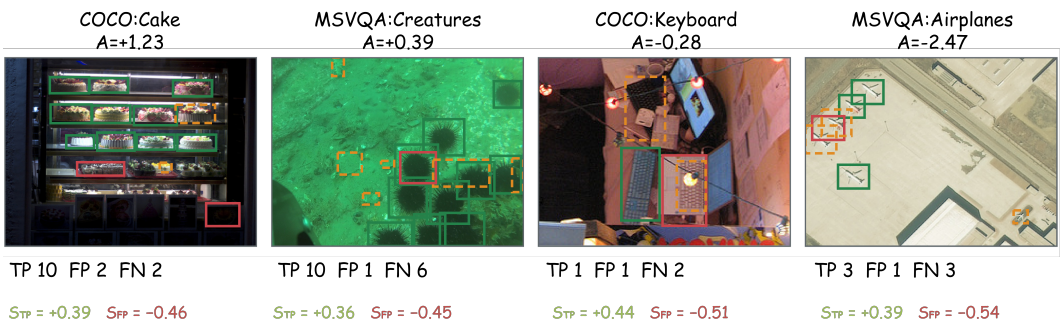

Figure A1: **Regional credit examples.** Four GRPO rollouts from COCO and MSVQA. Green, red, and dashed orange boxes denote TP, FP, and FN, respectively. A is the original response-level advantage; s_{TP} and s_{FP} summarize the normalized advantage changes for the two prediction groups.

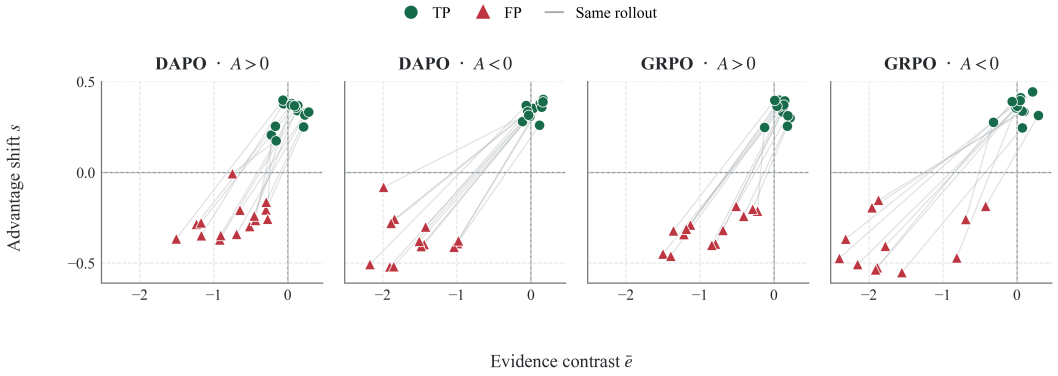

Figure A2: **TP–FP credit separation.** Each gray line connects the TP mean (green circle) and FP mean (red triangle) within one rollout. The axes show mean evidence contrast $\bar{e}_g$ and normalized advantage change s_g.

H Token-Level Credit

We examine whether evidence contrast distinguishes matched predictions (TP) from unmatched predictions (FP) within the same response. We replay rollouts from 10 source examples. For each rollout, we summarize the evidence contrast and advantage change over TP and FP box-coordinate spans separately:

$$\bar{e}_g = \frac{1}{|\mathcal{T}_g|} \sum_{t \in \mathcal{T}_g} e_t, \qquad s_g = \frac{1}{|\mathcal{T}_g|} \sum_{t \in \mathcal{T}_g} \frac{\widetilde{A}_t - A}{\alpha}, \quad g \in \{\mathrm{TP}, \mathrm{FP}\}. \tag{A2}$$

Here, $\mathcal{T}_g$ contains the tokens in the corresponding box-coordinate spans, including their delimiters. A positive s_g indicates an increase in advantage, while a negative value indicates a decrease. Figure A1 shows that correct and incorrect predictions coexist under both positive and negative response-level advantages. A shared outcome advantage therefore does not distinguish their regional quality. The TP mean shift exceeds the FP mean shift in all pairs. These results show that the credit signal distinguishes regional correctness within the inspected responses.