VECPO: VIEW-EQUIVARIANT CONSTRAINED POLICY OPTIMIZATION FOR ROBUST MULTI-VIEW REMOTE SENSING AGRICULTURAL REASONING

Anonymous authors
Paper under double-blind review

ABSTRACT

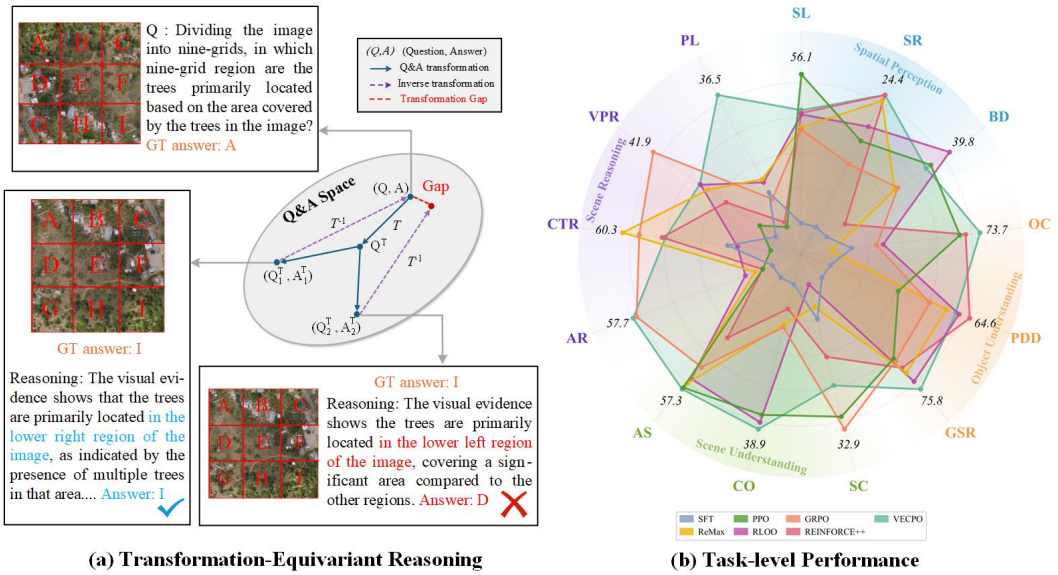
Agricultural multimodal large language models must reason over heterogeneous observations spanning ground-level, unmanned aerial vehicle, and satellite imagery. Although supervised fine-tuning on multi-view data improves domain knowledge, subsequent reinforcement learning faces two core challenges: (i) **transformation inconsistency**, where predictions become inconsistent under semantics-preserving spatial transformations; and (ii) **hidden group regression**, where aggregate optimization conceals regressions on underrepresented view–task groups. To address both challenges, we propose **View-Equivariant Constrained Policy Optimization (VECPO)**, a critic-free framework comprising Dual-Granularity Constrained Advantage (DGCA) and the Warm-Start Group Constraint Controller (WGCC). VECPO constructs an identity branch and eligible transformation branches with jointly transformed visual inputs and targets, canonicalizes predictions into a shared reference frame, and assigns correctness-gated equivariance rewards. DGCA couples a source-level relative advantage across all branches with branch-level relative advantages, while WGCC monitors view–task performance against supervised warm-start anchors and supplies group weights that constrain the dual-granularity advantage when a group deteriorates. Experiments on the 13-task AgroMind benchmark, including comparisons with PPO and multiple critic-free policy optimization baselines, show that VECPO achieves strong and competitive performance across diverse agricultural perception and reasoning tasks. Code is available in the supplementary material.

1 INTRODUCTION

Agriculture is central to global food security, climate resilience, and sustainable development (World Bank, 2026; Unicef et al., 2024; Alexopoulos et al., 2023; Zhai et al., 2020). Reliable interpretation of agricultural observations can support crop monitoring, disease diagnosis, yield assessment, and resource-efficient management. Large multimodal models (LMMs) offer a promising interface for these applications by combining visual evidence with language-based knowledge and reasoning (Yang et al., 2025b;a; Wang et al., 2024; Zhang et al., 2026a; Yin et al., 2024; Arshad et al., 2025; Ravindran et al., 2026; Yang et al., 2026; Li et al., 2025; Zhang et al., 2026b). This potential has motivated a growing range of agricultural datasets and benchmarks, including resources that span ground-level, unmanned aerial vehicle (UAV), and satellite imagery (Li et al., 2026; Zhang et al., 2026a; Gauba et al., 2026; Liu et al., 2024; Hughes et al., 2015; Wu et al., 2019).

Current progress in agricultural LMMs (Yang et al., 2025b; Wang et al., 2024; Liu et al., 2024; Yang et al., 2025a) is driven largely by supervised and reinforcement-based post-training. Supervised fine-tuning (SFT) injects domain knowledge and task formats through curated instruction data (Liu et al., 2023; Dai et al., 2023), while reinforcement learning with verifiable rewards further optimizes sampled responses (Shao et al., 2024; Guo et al., 2025; Ouyang et al., 2022). General actor–critic and critic-free methods provide different estimators and control variates for this process (Schulman et al., 2017; Li et al., 2023; Ahmadian et al., 2024; Hu et al., 2025; Shao et al., 2024). Recent multimodal variants further improve visual grounding, perception rewards, and reasoning-oriented optimization (Wang et al., 2026; Huang et al., 2026; Xiao et al., 2026), making policy optimization an increasingly important tool for domain-specific multimodal reasoning.

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.



(a) Transformation-Equivariant Reasoning **(b) Task-level Performance**

Figure 1: **(a)** Spatial transformations require predictions to change equivariantly and agree after inverse canonicalization. **(b)** Task-level results on AgroMind reveal heterogeneous gains and regressions across SFT and policy-optimization methods. VECPO addresses both challenges through correctness-gated equivariance learning and group-aware performance constraints.

However, applying these methods to heterogeneous agricultural imagery exposes two core challenges. First, *transformation inconsistency* arises when a semantics-preserving spatial transformation changes the prediction in an invalid way: flipping an image should transform a bounding box, grid location, or directional relation accordingly, while leaving a transformation-insensitive answer unchanged. Second, *hidden group regression* arises because aggregate reward can improve even as performance deteriorates on a particular ground-level, UAV-level, or satellite-level task. Standard rollout-level objectives do not explicitly encode either transformation correspondence or stability across fine-grained view–task groups.

Prior work offers useful but separate ideas for addressing these challenges. Equivariant architectures and consistency regularization encourage related inputs to produce compatible predictions (Cohen & Welling, 2016; Tarvainen & Valpola, 2017; Sohn et al., 2020), but generic augmentation does not specify how heterogeneous agricultural targets should transform, and agreement alone may reinforce consistently incorrect answers. Group DRO, sample reweighting, and adaptive policy optimization redistribute learning pressure across difficult or underrepresented groups (Sagawa et al., 2019; Qiao et al., 2025; Gao et al., 2026; Dai et al., 2026), but they do not jointly measure fine-grained view–task degradation relative to a supervised warm-start policy. These ideas motivate combining correctness-aware equivariance with explicit relative group constraints in a single policy-optimization framework.

Motivated by these observations, we propose VECPO, a critic-free framework with two complementary optimization modules. The first module, *Dual-Granularity Constrained Advantage* (DGCA), addresses transformation inconsistency. Specifically, VECPO constructs an identity branch and one or more transformed branches, maps transformation-sensitive targets, and uses canonicalized, correctness-gated consistency feedback. DGCA then couples a source-level relative advantage across all branches with branch-level relative advantages that retain within-prompt discrimination. The second module, *Warm-Start Group Constraint Controller* (WGCC), addresses hidden group regression by monitoring view–task performance against warm-start anchors and adaptively reweighting the dual-granularity advantage when a group deteriorates. This converts group-specific deficits into direct optimization pressure instead of assuming that aggregate reward improvements protect every subgroup. Together, DGCA and WGCC make both geometric correspondence and warm-start-relative performance retention explicit within a single critic-free objective. Our main contributions are summarized as follows:

- We propose VECPO, a critic-free policy optimization framework that introduces transformation-equivariant learning into multi-view agricultural reasoning. Its correctness-aware transformation branches supply DGCA, which combines source-level comparison across branches with branch-level discrimination within each prompt.
- We develop WGCC, which monitors view–task-group performance relative to warm-start anchors and adaptively constrains the dual-granularity advantage when a group deteriorates, making hidden group-level regressions an explicit optimization concern.
- Experiments on the 13-task AgroMind benchmark demonstrate that VECPO achieves strong and competitive performance across diverse agricultural perception and reasoning tasks, supporting the effectiveness of the overall framework.

2 Related Work

2.1 Multimodal Large Language Models for Agriculture

Multimodal large language models (MLLMs) are shifting agricultural remote sensing AI from isolated visual recognition toward knowledge-intensive diagnosis and reasoning (Zhu et al., 2025; Yang et al., 2025a; 2024; Yan et al., 2025). AgroGPT and Agri-LLaVA adapt general models through expert conversations and domain knowledge (Awais et al., 2025; Wang et al., 2024). Meanwhile, AgriBench, AgroBench, AgMMU, and Agri-CM^3 broaden evaluation from recognition to knowledge and management reasoning (Shinoda et al., 2025; Gauba et al., 2026; Wang et al., 2025; Zhou & Ryo, 2024). MIRAGE studies grounded expert consultation, AgroMind spans perception and reasoning across camera, UAV, and satellite imagery, and AgroOmni provides cross-scale instruction data (Dongre et al., 2025; Li et al., 2026; Zhang et al., 2026a).

However, existing agricultural MLLMs largely optimize and evaluate individual observations. They neither enforce task-dependent correspondence under safe spatial transformations nor protect individual view–task groups from hidden regression. VECPO addresses these gaps with canonicalized, correctness-gated transformation feedback and DGCA, together with warm-start-relative group constraints implemented by WGCC.

2.2 Reinforcement Learning for Multimodal Reasoning

Reinforcement learning has become a central post-training approach for improving language and multimodal reasoning (Schulman et al., 2017; Li et al., 2023; Ahmadian et al., 2024; Hu et al., 2025; Shao et al., 2024). PPO learns a value model for advantage estimation and limits policy drift through a clipped surrogate objective (Schulman et al., 2017); critic-free alternatives instead use greedy, leave-one-out, or group-relative baselines (Li et al., 2023; Ahmadian et al., 2024; Shao et al., 2024; Hu et al., 2025). Multimodal extensions further introduce difficulty-adaptive advantages, perception-aware optimization, visual-token dependence, and perception rewards (Gao et al., 2026; Wang et al., 2026; Huang et al., 2026; Xiao et al., 2026). These methods improve reward estimation, visual grounding, and reasoning-oriented optimization.

However, existing policy objectives do not jointly address the two failures exposed by heterogeneous multi-view reasoning: visual-aware rewards do not define transformation correspondence, and aggregate optimization can mask view–task regression. Group-robust objectives protect weak subpopulations (Sagawa et al., 2019), but do not encode task-dependent geometry. VECPO unifies both requirements: transformed rollouts and correctness gating supply DGCA, while WGCC constrains group behavior relative to the shared supervised warm start.

3 Method

3.1 Overview

VECPO is a critic-free post-training framework that makes transformation equivariance and group-aware stability explicit in policy optimization. As illustrated in Figure 2, it first constructs an identity branch and one or more safe transformation branches from each source example. The shared policy samples multiple responses from every branch, and their parsed answers are mapped to a common

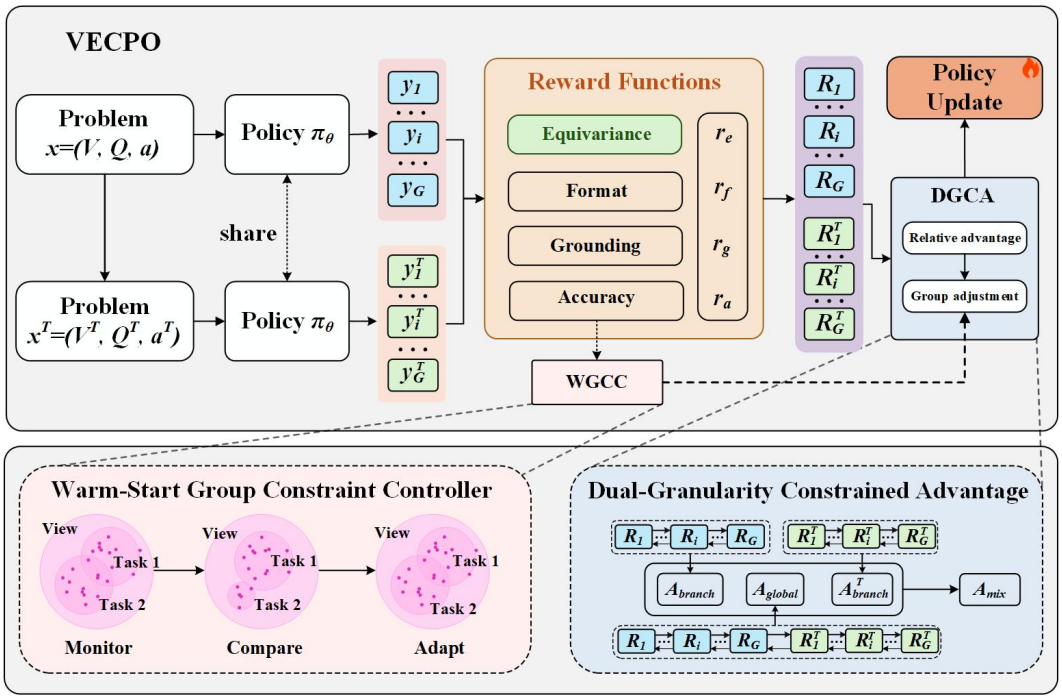


Figure 2: Overview of VECPO. A source and its safe transformation branches are processed by a shared policy. Parsed answers are canonicalized before computing the correctness-gated equivariance reward. DGCA couples source-level and branch-level relative advantages, while WGCC supplies adaptive constraint weights before the policy update.

reference frame. VECPO then introduces two complementary mechanisms. *Dual-Granularity Constrained Advantage* (DGCA) combines comparisons across transformation branches with discrimination among responses to the same branch. The *Warm-Start Group Constraint Controller* (WGCC) monitors view–task–answer-type groups against frozen warm-start anchors and upweights groups that deteriorate. Together, these mechanisms produce a group-adjusted advantage for the clipped policy update, without training a separate value model.

3.2 Transformation-Equivariant Rollouts

Let a source example be $z = (x, q, y, v, \tau, c)$, where x is the visual input, q is the question, y is the target, v is the acquisition view, τ is the task, and c is the answer type. The acquisition view identifies the imaging platform, such as ground, UAV, or satellite; a transformation acts within a view and never converts one acquisition view into another. Let $\mathcal{B}_z$ contain the identity branch and any number of candidate transformations. The indicator $m_{z,b} \in \{0, 1\}$ marks whether branch b admits an exact answer-space map. For an eligible branch,

$$z^b = \big(T_b^x(x), T_b^q(q), M_b(y), v, \tau, c\big), \qquad b \in \mathcal{B}_z, \quad m_{z,b} = 1, \tag{1}$$

where T_b^x transforms every image or frame consistently, T_b^q rewrites spatial language when required, and M_b transforms the target. The identity branch uses identity maps. In our implementation, candidate transformations are horizontal flip, vertical flip, and 180° rotation. Their effect is answer-type dependent: directional terms, grid cells, spatial options, and bounding boxes follow the image geometry, whereas invariant targets and image-option labels remain unchanged. Appendix B provides the complete mappings.

The rollout policy samples n_b responses from branch b, $\mathcal{O}_{z,b} = \{o_{b,k}\}_{k=1}^{n_b}$, and all responses derived from the source form $\mathcal{O}_z = \bigcup_{b \in \mathcal{B}_z} \mathcal{O}_{z,b}$. Because an equivariant answer may use different coordinates or labels after transformation, VECPO parses each response and maps its prediction back to

the source frame:

$$\widetilde{y}_{b,k} = C_b(\widehat{y}_{b,k}), \qquad C_b = M_b^{-1}. \tag{2}$$

Canonicalization is performed on task-level answers rather than free-form reasoning text. If an exact inverse map is unavailable, the transformation is omitted from equivariance comparisons. A fixed-shape implementation may retain a masked identity placeholder, but it receives no equivariance reward and no independent branch-level comparison. This rule prevents approximate or semantically invalid transformations from entering the learning signal. The general formulation permits arbitrary numbers of branches and responses; the reported configuration uses the identity branch, one sampled transformation branch, and two responses per branch.

3.3 Correctness-Gated Equivariance Reward

Each response receives the same task-aware base reward used by the matched policy optimization baselines, $r_i^{\text{base}} = a_i + 0.1f_i + 0.2h_i$. Here, a_i measures task accuracy, f_i checks the required response format, and h_i provides additional spatial grounding feedback for bounding-box tasks. The accuracy term is adapted to the answer type, using categorical correctness, relative numeric error, set overlap, or geometric overlap as appropriate. Complete definitions are given in Appendix C.

These rewards allow partial credit, which is useful for optimization but insufficient for determining whether cross-branch consistency should be rewarded. Therefore, VECPO computes a separate exactness indicator $u_i = E_c(\widetilde{y}_i, y)$. For response i, let $\mathcal{C}_i = \{j \in \mathcal{O}_z : b(j) \neq b(i), m_{z,b(i)}m_{z,b(j)} = 1\}$ contain responses from other eligible branches. Its equivariance score and final reward are

$$\begin{aligned} c_i &= \begin{cases} \dfrac{1}{|\mathcal{C}_i|}\displaystyle\sum_{j\in\mathcal{C}_i} u_i u_j E_c(\widetilde{y}_i, \widetilde{y}_j), & \mathcal{C}_i \neq \varnothing, \\ 0, & \mathcal{C}_i = \varnothing, \end{cases} \\ r_i &= r_i^{\text{base}} + \alpha_t c_i. \end{aligned} \tag{3}$$

The product $u_i u_j$ opens the gate only when both predictions are correct; cross-branch agreement therefore receives no bonus when the answers are wrong. The coefficient α_t is warmed up gradually so that consistency does not dominate before the policy produces reliable answers. This reward encourages the intended change under a transformation, rather than invariance of raw answer strings.

3.4 Dual-Granularity Constrained Advantage

Source-level normalization compares all responses derived from the same source and therefore captures reward differences across transformation branches. Branch-level normalization, by contrast, ranks responses generated from the identical prompt and preserves fine-grained within-prompt discrimination. Each signal alone is limited: the larger source-level set can dilute rankings among responses to the same prompt, while branch-level normalization cannot reveal whether one transformed branch is systematically harder than another. DGCA retains both by mixing the two advantages. For a response set S, define $\mathrm{StdAdv}(i; S) = (r_i - \mu_S)/(\sigma_S + \epsilon_A)$. We use $\chi_i = \mathbf{1}[m_{z,b(i)} = 1 \wedge |\mathcal{O}_{z,b(i)}| > 1]$ to indicate that response i belongs to an eligible branch with at least two responses. The two advantages and their mixture are

$$\begin{aligned} A_i^{\text{src}} &= \mathrm{StdAdv}(i; \mathcal{O}_z), \\ A_i^{\text{br}} &= \begin{cases} \mathrm{StdAdv}(i; \mathcal{O}_{z,b(i)}), & \chi_i = 1, \\ A_i^{\text{src}}, & \text{otherwise}, \end{cases} \\ A_i^{\text{DG}} &= (1-\rho)A_i^{\text{src}} + \rho A_i^{\text{br}}. \end{aligned} \tag{4}$$

The fallback avoids a degenerate local normalization for a singleton or masked branch. The coefficient ρ controls the balance between transformation-aware source comparison and within-prompt discrimination. We set $\rho = 1/|B_z|$ (equal to 0.5 under the reported two-branch configuration).

3.5 Warm-Start Group Constraint Controller

Equivariance within a source does not prevent aggregate improvements from masking deterioration in a particular acquisition view, task, or answer type. WGCC therefore maintains controller state

for each group $g = (v, \tau, c)$. Let $\widehat{u}_g^{(t)}$ be the mean exact correctness observed for group g at step t. During an initial warm-start phase, the controller estimates a group-specific anchor b_g and then freezes it. Subsequent updates smooth the observed correctness and perform a projected multiplier step:

$$\begin{aligned} \overline{u}_g^{(t)} &= \gamma \overline{u}_g^{(t-1)} + (1-\gamma)\widehat{u}_g^{(t)}, \\ \lambda_g &\leftarrow \mathrm{clip}(\lambda_g + \eta_\lambda(b_g - \overline{u}_g - \delta), 0, \lambda_{\max}) \,. \end{aligned} \tag{5}$$

Thus, persistent degradation beyond the tolerance δ increases λ_g, while fluctuations inside the tolerance do not. Groups absent from a batch retain their previous state. Because the anchors are constructed from training-time warm-start statistics, the controller does not use the evaluation set to define its constraints.

WGCC converts the multipliers into normalized, bounded response weights and applies them to the dual-granularity advantage:

$$\begin{aligned} \widetilde{w}_i &= 1 + \lambda_{g(i)}, \\ \overline{w} &= \frac{1}{B}\sum_{k=1}^{B} \widetilde{w}_k, \\ w_i &= \mathrm{clip}\left(\frac{\widetilde{w}_i}{\overline{w}}, w_{\min}, w_{\max}\right), \\ \widehat{A}_i^{\mathrm{DGCA}} &= w_i A_i^{\mathrm{DG}}. \end{aligned} \tag{6}$$

Batch-mean normalization preserves the overall advantage scale, while clipping prevents one group from dominating an update. WGCC consequently reallocates learning pressure toward deteriorating groups rather than simply enlarging the entire policy gradient. Controller schedules and bounds are reported in Appendix D.

3.6 Policy Objective

Let $\varrho_{i,t}(\theta)$ denote the token-level importance ratio between the current and rollout policies. VECPO applies the response-level constrained advantage to every valid generated token and maximizes

$$\begin{aligned} \mathcal{J}_{\mathrm{VECPO}}(\theta) = \mathbb{E}_{i,t}\Big[\min\big(&\varrho_{i,t}(\theta)\widehat{A}_i^{\mathrm{DGCA}}, \\ &\mathrm{clip}(\varrho_{i,t}(\theta), 1-\varepsilon, 1+\varepsilon)\widehat{A}_i^{\mathrm{DGCA}}\big)\Big] - \beta \mathcal{D}_{\mathrm{KL}}(\pi_\theta \| \pi_{\mathrm{ref}}). \end{aligned} \tag{7}$$

The KL term regularizes the policy toward the shared SFT reference. VECPO therefore preserves a critic-free update while adding transformation-aware credit assignment and group-relative protection. Appendix D provides the complete training configuration and pseudocode.

4 Experiments

4.1 Experiments settings

Models, Data, and Baselines. We use Qwen2.5-VL-7B-Instruct (Bai et al., 2025) as the base model and first adapt it to the agricultural domain through supervised fine-tuning on AgroOmni (Zhang et al., 2026a). The training data follow an entity-disjoint protocol: all records associated with the same primary entity are assigned exclusively to either SFT or RL, and the RL partition contains 49,152 source prompts. Every policy-optimization run starts from the same SFT checkpoint. We compare VECPO with the base model, the SFT checkpoint, the actor–critic PPO baseline, and four critic-free baselines: GRPO, ReMax, RLOO and Reinforce++(Schulman et al., 2017; Shao et al., 2024; Li et al., 2023; Ahmadian et al., 2024; Hu et al., 2025). The RL comparisons use the same source partition, task-aware reward implementation, and fixed update horizon, while retaining each estimator's intrinsic sampling procedure. Additional split construction and prompt details are provided in Appendix A.

Table 1: Preliminary performance comparison on AgroMind using training seed 42. We report accuracy (%) for each task and the 13-task macro average. All RL methods start from the same SFT checkpoint. Light red and light blue cells denote the best and second-best result in each column, respectively.

Method	Spatial Perception			Object Understanding			Scene Understanding			Scene Reasoning				Overall
	SL	SR	BD	OC	PDD	GSR	SC	CO	AS	AR	CTR	VPR	PL	
Base model	32.76	20.92	28.19	69.60	36.22	40.58	31.02	20.73	34.10	23.73	56.57	18.74	33.66	34.37
SFT Baseline	30.75	20.84	29.34	70.43	47.35	40.33	30.78	20.99	41.06	23.47	57.32	19.42	34.00	35.85
+PPO	56.14	23.17	38.19	73.22	55.74	66.08	32.65	37.11	57.08	25.84	56.10	22.38	33.11	44.37
+GRPO	46.71	22.55	35.38	71.07	59.61	66.83	32.89	26.28	54.09	56.87	59.83	41.92	33.83	46.76
+ReMax	47.10	24.26	35.11	69.92	61.70	70.67	30.54	25.95	56.49	27.74	60.30	31.90	34.34	44.31
+RLOO	49.31	23.56	39.78	71.24	63.26	73.42	30.11	38.03	55.79	30.11	57.04	33.21	34.26	46.09
+REINFORCE++	49.56	24.42	30.88	73.37	64.57	69.08	31.50	24.01	49.81	25.90	59.18	28.55	33.23	43.39
+VECPO	50.03	24.42	37.80	73.74	62.87	75.83	32.05	38.88	57.30	57.73	59.09	33.38	36.52	49.20

Training Details. The reported RL runs use training seed 42. The common actor learning rate is 10^{-6}, and prompt and response lengths are capped at 4,096 and 512 tokens, respectively. Where the estimator permits, each update uses a matched budget of 256 generated responses. VECPO realizes this budget with 64 sources, two branches per source, and two responses per branch; ReMax (Li et al., 2023) retains its stochastic-response/greedy-baseline construction, and PPO (Schulman et al., 2017) retains its learned critic. VECPO is trained on four A100 GPUs with rollout temperature 1.0 and top-p 0.99. More training details are provided in Appendix D.

Evaluation Benchmark and Metrics. We evaluate all models on AgroMind (Li et al., 2026), which contains 28,482 examples across 13 agricultural tasks grouped into spatial perception, object understanding, scene understanding, and scene reasoning. All models use the same fixed evaluation pipeline. We report task-level accuracy and the unweighted macro average over the 13 tasks as the primary aggregate. More details are provided in Appendix E.

4.2 Main Results

We compare VECPO with the base model, SFT, and five representative policy- optimization methods in Table 1. VECPO achieves the strongest overall performance, with a 13-task macro accuracy of 49.20. It exceeds all five policy-optimization baselines by 2.44–5.81 points and improves the SFT baseline by 13.35 points. Beyond the aggregate score, VECPO ranks among the top two methods on nine tasks, including six outright leads and one tied lead, with leading results spanning all four AgroMind capability families. On the remaining four tasks, its gap to the best method is at most 1.98 points, indicating a balanced performance profile rather than specialization to a small subset of tasks. VECPO also improves every task over the supervised warm start, so its aggregate gain does not mask a task-level regression. Together, these results establish broad task-level effectiveness and support VECPO's goal of mitigating hidden regression while preserving strong perception and reasoning performance.

4.3 Ablation Study

Component ablations. Table 2 evaluates the contribution of DGCA and WGCC by removing each component from VECPO. Compared with w/o WGCC, the complete method improves nine of the 13 tasks, with the largest margins appearing on SL (8.04 points), GSR (7.41), VPR (5.21), and CO (4.25). The improvements cover all four capability families, supporting WGCC's effectiveness in protecting heterogeneous view–task groups. Removing DGCA similarly degrades eight tasks. Its largest effects appear on AR and AS, where the complete method gains 30.38 and 10.01 points, respectively, while improvements also extend to spatial perception and object understanding. Together, these results show that DGCA and WGCC make complementary contributions to VECPO's task-level performance.

Table 2: Ablation study of DGCA and WGCC on AgroMind. We report accuracy (%) across all 13 tasks. Best results per column are highlighted in light red.

Method	Spatial Perception			Object Understanding			Scene Understanding			Scene Reasoning			
	SL	SR	BD	OC	PDD	GSR	SC	CO	AS	AR	CTR	VPR	PL
SFT Baseline	30.75	20.84	29.34	70.43	47.35	40.33	30.78	20.99	41.06	23.47	57.32	19.42	34.00
+GRPO	46.71	22.55	35.38	71.07	59.61	66.83	32.89	26.28	54.09	56.87	59.83	41.92	33.83
VECPO	50.03	24.42	37.80	73.74	62.87	75.83	32.05	38.88	57.30	57.73	59.09	33.38	36.52
Component Ablation													
w/o DGCA	53.09	22.78	38.02	73.02	63.78	73.58	33.80	34.48	47.29	27.35	60.11	30.54	34.13
w/o WGCC	41.99	24.73	37.86	70.66	60.22	68.42	33.55	34.63	54.41	54.96	60.95	28.17	34.64

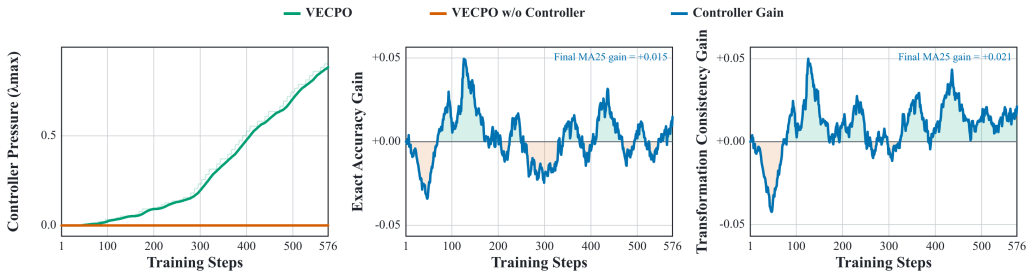


Figure 3: Training dynamics of the matched controller ablation. (a) Maximum controller pressure. (b–c) The 25-step trailing-mean gains of VECPO over VECPO w/o Controller in exact accuracy and view consistency, ending at $+0.015$ and $+0.021$, respectively.

5 Further Discussion

5.1 Transformation accuracy.

Figure 4 directly examines the behavior targeted by view-equivariant optimization. For every eligible example, we apply the same non-identity transformation to the image and any transformation-sensitive target, map the prediction back to the original reference frame, and then compute task accuracy. The balanced diagnostic contains 256 examples from each of 13 tasks, or 3328 examples per transformation for each model. VECPO achieves transformed accuracies of 50.9%, 50.1%, and 50.8% under horizontal flip, vertical flip, and 180° rotation, compared with 47.1%, 46.5%, and 46.3% for GRPO (Shao et al., 2024). The corresponding gains are 3.8, 3.6, and 4.4 percentage points, yielding an average improvement of 3.94 points across the three transformations. The consistent point-estimate gains support the intended role of paired transformation training: VECPO more often preserves task correctness after the visual input and its spatial semantics change.

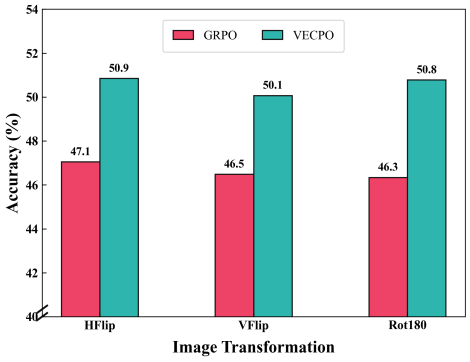


Figure 4: Canonicalized transformed accuracy of GRPO and VECPO under horizontal flip, vertical flip, and 180° rotation on the post-hoc 13-task diagnostic subset.

5.2 Controller effectiveness.

As Figure 3 shows, after warm-up VECPO steadily increases controller pressure as underperforming groups appear, reaching a maximum multiplier of 0.921. In contrast, the ablated model keeps all multipliers at zero and all group weights at one. VECPO also attains higher mean training exact accuracy (0.5557 vs. 0.5522) and view consistency (0.4909 vs. 0.4797), with final 25-step moving-

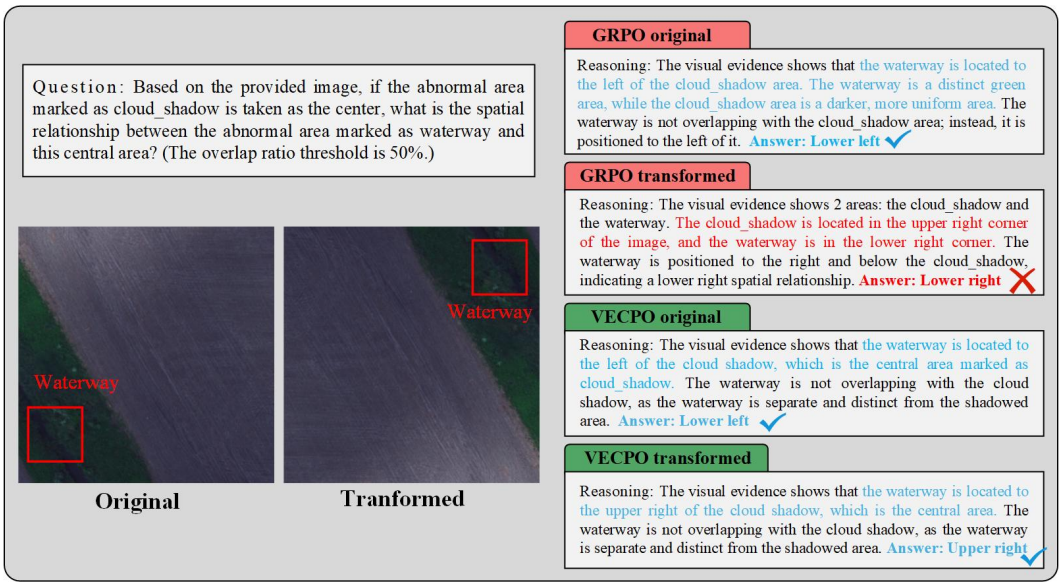


Figure 5: Qualitative comparison on a transformation-sensitive spatial- relation example. Both GRPO and VECPO answer the original image correctly. After the spatial transformation, GRPO produces an inconsistent relation, whereas VECPO changes the answer from lower left to upper right and matches the transformed target.

average gains of 0.015 and 0.021, respectively. Although the exact-accuracy gain fluctuates during training, the later consistency gain remains more persistently positive. This indicates that the controller actively reallocates optimization pressure, primarily improving cross-view stability while also yielding a modest correctness benefit.

5.3 Qualitative Analysis

Figure 5 illustrates the failure mode targeted by view-equivariant optimization. In the original image, the waterway lies to the lower left of the cloud-shadow region, which both methods identify correctly. The transformation moves the same region to the upper right, so a correct prediction must change with the visual geometry rather than repeat or arbitrarily alter the original relation. GRPO (Shao et al., 2024) instead predicts lower right, despite being correct on the original view, revealing a transformation-induced reasoning inconsistency. VECPO preserves correctness across the pair: it updates both its spatial explanation and final answer to upper right. This example provides an interpretable instance of the behavior summarized by the transformed-accuracy results in Figure 4; it is illustrative rather than a standalone aggregate comparison.

6 Conclusion

We introduce VECPO, a critic-free framework combining correctness-gated equivariant rollouts with Dual-Granularity Constrained Advantage (DGCA) and a Warm-Start Group Constraint Controller (WGCC). DGCA couples source-level comparison across transformation branches with branch-level discrimination, while WGCC upweights groups that deteriorate relative to warm-start anchors. On the 13-task AgroMind benchmark, VECPO achieves competitive aggregate performance; diagnostic analyses further show higher canonicalized accuracy than GRPO under all three evaluated transformations and improved training-time cross-view consistency with WGCC. These results support making geometric correspondence and group-aware retention explicit in multimodal policy optimization. Future work will evaluate multiple seeds and datasets, broader safe transformations, and stronger group-level no-regression guarantees.

AI USE STATEMENT

Generative AI tools were used to assist with drafting and polishing the manuscript, checking consistency, and preparing LaTeX, as well as literature retrieval and discovery to identify potentially relevant related work. AI outputs were not treated as scientific evidence or cited as sources. The authors retain responsibility for verifying the cited literature and for the accuracy and integrity of the manuscript, including its scientific claims, numerical results, and references.

ETHICS STATEMENT

Our research is conducted within the domain of agricultural multimodal reasoning. We use existing academic datasets, collect no new data, and involve neither human subjects nor autonomous field interventions. Our goal is to improve transformation-consistent and group-stable reasoning across agricultural perception and reasoning tasks. The underlying imagery and annotations may nevertheless reflect geographic, sensor, crop, and task imbalances. Such uses require domain-expert verification and application-specific validation. Redistribution and deployment should also respect the licenses, provenance, privacy requirements, and permitted uses of the underlying datasets.

REPRODUCIBILITY STATEMENT

To support reproducibility, we document the proposed method and the available evaluation provenance throughout the paper and appendices. Section 3 specifies the transformation rollouts, reward, DGCA, WGCC, and policy objective, while Section 4 describes the common training and evaluation protocol. Appendix A documents the deterministic entity-disjoint data partition and prompt construction; Appendix B gives the transformation and canonicalization rules; Appendix C defines the complete reward functions; and Appendix D reports the SFT and RL configurations, random seed, rollout budget, controller settings, baseline implementations, and training pseudocode.

REFERENCES

Arash Ahmadian, Chris Cremer, Matthias Gallé, Marzieh Fadaee, Julia Kreutzer, Olivier Pietquin, Ahmet Üstün, and Sara Hooker. Back to basics: Revisiting reinforce-style optimization for learning from human feedback in llms. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 12248–12267, 2024.

Angelos Alexopoulos, Konstantinos Koutras, Sihem Ben Ali, Stefano Puccio, Alessandro Carella, Roberta Ottaviano, and Athanasios Kalogeras. Complementary use of ground-based proximal sensing and airborne/spaceborne remote sensing techniques in precision agriculture: A systematic review. *Agronomy*, 13(7):1942, 2023.

Muhammad Arbab Arshad, Talukder Zaki Jubery, Tirtho Roy, Rim Nassiri, Asheesh K Singh, Arti Singh, Chinmay Hegde, Baskar Ganapathysubramanian, Aditya Balu, Adarsh Krishnamurthy, et al. Leveraging vision language models for specialized agricultural tasks. In *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 6320–6329. IEEE, 2025.

Muhammad Awais, Ali Husain Salem Abdulla Alharthi, Amandeep Kumar, Hisham Cholakkal, and Rao Muhammad Anwer. Agrogpt: Efficient agricultural vision-language model with expert tuning. In *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 5687–5696. IEEE, 2025.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-vl technical report, 2025. URL https://arxiv.org/abs/2502.13923.

Taco Cohen and Max Welling. Group equivariant convolutional networks. In *International conference on machine learning*, pp. 2990–2999. PMLR, 2016.

David Dai, Peilin Chen, Chanakya Ekbote, and Paul Liang. Qoq-med: Building multimodal clinical foundation models with domain-aware grpo training. *Advances in Neural Information Processing Systems*, 38:37406–37453, 2026.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale N Fung, and Steven Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning. *Advances in neural information processing systems*, 36:49250–49267, 2023.

Vardhan Dongre, Chi Gui, Shubham Garg, Hooshang Nayyeri, Gokhan Tur, Dilek Hakkani-Tür, and Vikram S Adve. Mirage: A benchmark for multimodal information-seeking and reasoning in agricultural expert-guided conversations. *arXiv preprint arXiv:2506.20100*, 2025.

Haowen Gao, Zhenyu Zhang, Liang Pang, Fangda Guo, Hongjian Dou, Guannan Lv, Shaoguo Liu, Tingting Gao, Huawei Shen, and Xueqi Cheng. Diva-grpo: Enhancing multimodal reasoning through difficulty-adaptive variant advantage. *arXiv preprint arXiv:2603.01106*, 2026.

Aruna Gauba, Irene Pi, Yunze Man, Ziqi Pang, Vikram Adve, and Yu-Xiong Wang. Agmmu: A comprehensive agricultural multimodal understanding benchmark. *Advances in Neural Information Processing Systems*, 38, 2026.

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

Jian Hu, Jason Klein Liu, Haotian Xu, and Wei Shen. Reinforce++: Stabilizing critic-free policy optimization with global advantage normalization. *arXiv preprint arXiv:2501.03262*, 2025.

Siyuan Huang, Xiaoye Qu, Yafu Li, Yun Luo, Zefeng He, Daizong Liu, and Yu Cheng. Spotlight on token perception for multimodal reinforcement learning. In *International Conference on Learning Representations*, volume 2026, pp. 103291–103325, 2026.

David Hughes, Marcel Salathé, et al. An open access repository of images on plant health to enable the development of mobile disease diagnostics. *arXiv preprint arXiv:1511.08060*, 2015.

Qingmei Li, Yang Zhang, Zurong Mai, Yuhang Chen, Henglian Huang, Jiarui Zhang, Zhiwei Zhang, Yibin Wen, Weijia Li, Haohuan Fu, et al. Can large multimodal models understand agricultural scenes? benchmarking with agromind. *Advances in Neural Information Processing Systems*, 38, 2026.

Wenyuan Li, Shunlin Liang, Keyan Chen, Yongzhe Chen, Han Ma, Jianglei Xu, Yichuan Ma, Shikang Guan, Husheng Fang, and Zhenwei Shi. Agrifm: A multi-source temporal remote sensing foundation model for crop mapping. *arXiv e-prints*, pp. arXiv–2505, 2025.

Ziniu Li, Tian Xu, Yushun Zhang, Zhihang Lin, Yang Yu, Ruoyu Sun, and Zhi-Quan Luo. Remax: A simple, effective, and efficient reinforcement learning method for aligning large language models. *arXiv preprint arXiv:2310.10505*, 2023.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023.

Xiang Liu, Zhaoxiang Liu, Huan Hu, Zezhou Chen, Kohou Wang, Kai Wang, and Shiguo Lian. A multimodal benchmark dataset and model for crop disease diagnosis. In *European conference on computer vision*, pp. 157–170. Springer, 2024.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35: 27730–27744, 2022.

Rui Qiao, Zhaoxuan Wu, Jingtan Wang, Pang Wei Koh, and Bryan Kian Hsiang Low. Group-robust sample reweighting for subpopulation shifts via influence functions. In *International Conference on Learning Representations*, volume 2025, pp. 36959–36980, 2025.

Dinesh Jackson Samuel Ravindran, Inna Skarga-Bandurova, Sivakumar V, Muhammad Awais, and Mithra S. Agrollm: Connecting farmers and agricultural practices through large language models for enhanced knowledge transfer and practical application. *AgriEngineering*, 8(1):38, 2026.

Shiori Sagawa, Pang Wei Koh, Tatsunori B Hashimoto, and Percy Liang. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. *arXiv preprint arXiv:1911.08731*, 2019.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

Risa Shinoda, Nakamasa Inoue, Hirokatsu Kataoka, Masaki Onishi, and Yoshitaka Ushiku. Agrobench: Vision-language model benchmark in agriculture. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 7634–7644. IEEE, 2025.

Kihyuk Sohn, David Berthelot, Nicholas Carlini, Zizhao Zhang, Han Zhang, Colin Raffel, Ekin Dogus Cubuk, Alexey Kurakin, and Chun-Liang Li. Fixmatch: Simplifying semi-supervised learning with consistency and confidence. *Advances in neural information processing systems*, 33: 596–608, 2020.

Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. *Advances in neural information processing systems*, 30, 2017.

Unicef et al. The state of food security and nutrition in the world 2024. 2024.

Haotian Wang, Yi Guan, Fanshu Meng, Chao Zhao, Lian Yan, Yang Yang, and Jingchi Jiang. Agricm3: A chinese massive multi-modal, multi-level benchmark for agricultural understanding and reasoning. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 11729–11754, 2025.

Liqiong Wang, Teng Jin, Jinyu Yang, Ales Leonardis, Fangyi Wang, and Feng Zheng. Agri-llava: Knowledge-infused large multimodal assistant on agricultural pests and diseases. *arXiv preprint arXiv:2412.02158*, 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.

World Bank. Agriculture and food: What we do. `https://www.worldbank.org/en/topic/agriculture/brief/agriculture-and-food-what-we-do`, 2026.

Xiaoping Wu, Chi Zhan, Yu-Kun Lai, Ming-Ming Cheng, and Jufeng Yang. Ip102: A large-scale benchmark dataset for insect pest recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 8787–8796, 2019.

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.

Jian Yan, Xingfa Gu, and Yuxing Chen. Cropsts: A remote sensing foundation model for cropland classification with decoupled spatiotemporal attention. *Remote Sensing*, 17(14):2481, 2025.

Bo Yang, Yunkui Chen, Lanfei Feng, Yu Zhang, Xiao Xu, Jianyu Zhang, Nueraili Aierken, Runhe Huang, Hongjian Lin, Yibin Ying, et al. Agrigpt-vl: Agricultural vision-language understanding suite. *arXiv preprint arXiv:2510.04002*, 2025a.

Bo Yang, Yu Zhang, Lanfei Feng, Yunkui Chen, Jianyu Zhang, Xiao Xu, Nueraili Aierken, Yurui Li, Yuxuan Chen, Guijun Yang, et al. Agrigpt: A large language model ecosystem for agriculture. *arXiv preprint arXiv:2508.08632*, 2025b.

Bo Yang, Lanfei Feng, Yunkui Chen, Yu Zhang, Jianyu Zhang, Xiao Xu, Nueraili Aierken, and Shijian Li. Agrigpt-omni: A unified speech–vision–text framework for multilingual agricultural intelligence. In *Proceedings of the ACM Web Conference 2026*, pp. 9800–9810, 2026.

Shuting Yang, Zehui Liu, Wolfgang Mayer, Ningpei Ding, Ying Wang, Yu Huang, Pengfei Wu, Wanli Li, Lin Li, Hong-Yu Zhang, et al. Shizishangpt: An agricultural large language model integrating tools and resources. In *International Conference on Web Information Systems Engineering*, pp. 284–298. Springer, 2024.

Shukang Yin, Chaoyou Fu, Sirui Zhao, Ke Li, Xing Sun, Tong Xu, and Enhong Chen. A survey on multimodal large language models. *National Science Review*, 11(12):nwae403, 2024.

Zhaoyu Zhai, José Fernán Martínez, Victoria Beltran, and Néstor Lucas Martínez. Decision support systems for agriculture 4.0: Survey and challenges. *Computers and Electronics in Agriculture*, 170:105256, 2020.

Jiarui Zhang, Junqi Hu, Zurong Mai, Yang Liu, Yuhang Chen, Shuohong Lou, Henglian Huang, Hong Cheng, Lingyuan Zhao, Jianxi Huang, et al. Agroomni: A large-scale multi-view agricultural dataset for cross-scale multimodal reasoning. *arXiv preprint arXiv:2603.14342*, 2026a.

Zhixing Zhang, Jesen Zhang, Hao Liu, Qinhan Lv, Jing Yang, Kaitong Cai, and Keze Wang. Agriworld: A world tools protocol framework for verifiable agricultural reasoning with code-executing llm agents. *arXiv preprint arXiv:2602.15325*, 2026b.

Yutong Zhou and Masahiro Ryo. Agribench: A hierarchical agriculture benchmark for multimodal large language models. In *European Conference on Computer Vision*, pp. 207–223. Springer, 2024.

Hongyan Zhu, Shuai Qin, Min Su, Chengzhi Lin, Anjie Li, and Junfeng Gao. Harnessing large vision and language models in agriculture: A review. *Frontiers in Plant Science*, 16:1579355, 2025.

APPENDIX CONTENTS

A Data Protocol and Prompt Construction **14**

A.1 AgroOmni Training Corpus 14

A.2 Entity-Disjoint SFT/RL Partition 14

A.3 Stratified Construction of the RL Partition 15

A.4 Training Prompt and SFT Target 16

B Transformation Rules and Answer Canonicalization **16**

B.1 Acquisition Views and Paired Transformations 16

B.2 Transformation Rules by Target Type 17

C Reward Functions **18**

C.1 Task-Aware Base Reward 18

C.2 Correctness-Gated Equivariance Reward 18

D Training Configuration and Procedure **19**

D.1 SFT Training Configuration 19

D.2 RL Training Configuration 20

D.3 VECPO Training Pseudocode 20

D.4 Reinforcement-Learning Baselines 21

E AgroMind Benchmark and Qualitative Cases **22**

E.1 AgroMind Benchmark 22

E.2 Qualitative Cases 23

A DATA PROTOCOL AND PROMPT CONSTRUCTION

A.1 AGROOMNI TRAINING CORPUS

AgroOmni (Zhang et al., 2026a) is a large-scale agricultural multimodal corpus that combines observations from ground cameras, unmanned aerial vehicles (UAVs), and satellites. It contains 276,709 question–answer pairs, organized into 262,432 training records and 14,277 records in an independent development split. The training pool contains 126,447 primary entities, while the development split contains 6,737 disjoint entities. Both splits cover 56 view-conditioned level-3 task categories and seven answer types. Table 3 reports the exact composition by acquisition view.

Each normalized record contains a unique `sample_id`, a primary `entity_id`, the acquisition view, level-2 and level-3 task labels, the answer type, one or more image references, the question, and the gold answer; multiple-choice options are retained when present. The seven answer types comprise single choice, Boolean, count, multi-select, short text, numeric, and bounding box. Their exact distributions in the training pool and its two partitions are given in Table 4. The data builder validates required fields, unique sample identifiers, and image-path safety. A deterministic 2,048-example monitor is derived from the independent development split for both SFT and RL.

A.2 ENTITY-DISJOINT SFT/RL PARTITION

We partition the normalized AgroOmni (Zhang et al., 2026a) training pool at the level of primary entities rather than individual question–answer pairs. Let $\mathcal{D}_e$ denote all records associated with

entity e. The protocol selects a set of entities $\mathcal{E}_{\mathrm{RL}}$ and defines

$$\mathcal{D}_{\mathrm{RL}} = \bigcup_{e \in \mathcal{E}_{\mathrm{RL}}} \mathcal{D}_e, \qquad \mathcal{D}_{\mathrm{SFT}} = \bigcup_{e \notin \mathcal{E}_{\mathrm{RL}}} \mathcal{D}_e, \tag{8}$$

subject to $|\mathcal{D}_{\mathrm{RL}}| = 49{,}152$. Consequently, every record belonging to a primary entity is assigned wholly to SFT or wholly to RL, and

$$\mathcal{E}_{\mathrm{SFT}} \cap \mathcal{E}_{\mathrm{RL}} = \emptyset. \tag{9}$$

All remaining training records form the SFT partition. This yields 213,280 SFT records and 49,152 RL records, corresponding to 81.27% and 18.73% of the training pool, respectively.

The partition is deterministic under seed 42. Before selection, records are grouped by **entity_id**, and each entity is represented by the counts of its records in the stratification cells described below. The initial selection adds complete entities toward the 49,152-record RL target. The protocol then exchanges entities of equal record count when necessary so that the total size does not change. These exchanges enforce two coverage conditions: every view–level-3 task represented in the training pool must occur in both SFT and RL, and every view–level-2–answer-type stratum must retain records in both partitions. The final checks verify row conservation and zero sample- or entity-level overlap between SFT and RL. Reused auxiliary option images are reported separately by the split audit and do not alter the entity-disjoint assignment rule.

A.3 Stratified Construction of the RL Partition

The RL partition is designed to preserve the natural AgroOmni (Zhang et al., 2026a) distribution while increasing the representation of rare and structurally demanding examples. We define each stratum as

$$s = (v, t, a), \tag{10}$$

where v is the acquisition view, t is the level-2 task, and a is the answer type. Let n_s denote the number of available training records in stratum s, and let n_s^{str} denote the number of structured records in that stratum. A record is treated as structured if it has a bounding-box, count, or numeric answer; contains multiple images; or belongs to one of the configured spatially structured view–level-3 tasks.

The continuous target allocation for stratum s is

$$q_s = 49{,}152 \left(0.5 \frac{n_s}{\sum_k n_k} + 0.3 \frac{\sqrt{n_s}}{\sum_k \sqrt{n_k}} + 0.2 \frac{\sqrt{n_s^{\mathrm{str}}}}{\sum_k \sqrt{n_k^{\mathrm{str}}}} \right). \tag{11}$$

The three terms play complementary roles. The natural-distribution term, weighted by 0.5, preserves the empirical frequency of common view–task–answer strata. The square-root-frequency term, weighted by 0.3, increases the relative allocation of less frequent strata without forcing a uniform distribution. The remaining 0.2 square-root structured-frequency term emphasizes geometry, counting, numeric, and multi-image reasoning examples.

The continuous targets are converted to capacity-constrained integer quotas with the largest-remainder method. Each stratum receives at least one RL record while retaining at least one record for SFT whenever the corpus capacity permits. Because entity disjointness is the primary constraint, the protocol never fills these quotas by independently moving individual rows. Instead, it computes a deterministic selection propensity for each entity from the mixture above, ranks entities using the fixed seed, and adds complete entities until exactly 49,152 records have been selected. Equal-size entity swaps are then used to restore missing level-3 task coverage and to minimize the normalized squared deviation between the realized and target stratum counts. Thus, the final RL distribution approximates the mixture targets while preserving exact size, complete entities, and task coverage in both partitions.

Finally, a 2,048-example RL probe is drawn from the frozen RL partition for runtime diagnosis. Its allocation follows square-root stratum frequency and cycles through the available view–level-3 tasks within each stratum. The probe is a subset of the RL data, not an additional training partition, and does not replace the full 49,152-example policy-optimization set.

Table 3: Statistics of AgroOmni and the resulting training partitions. SFT and RL are disjoint subsets of the training pool; the development split is independent of both.

Data subset	Records	Entities	Ground	UAV	Satellite
Full corpus	276,709	133,184	193,265	20,181	63,263
Training pool	262,432	126,447	183,217	19,214	60,001
Independent split	14,277	6,737	10,048	967	3,262
SFT partition	213,280	100,273	152,974	14,225	46,081
RL partition	49,152	26,174	30,243	4,989	13,920

Table 4: Answer-type distribution in the AgroOmni training pool and the entity-disjoint SFT/RL partitions.

Answer type	Training pool	SFT	RL
Single choice	151,732	126,062	25,670
Boolean	47,421	38,926	8,495
Count	31,766	24,865	6,901
Multi-select	14,511	11,063	3,448
Short text	8,699	6,841	1,858
Numeric	5,662	4,103	1,559
Bounding box	2,641	1,420	1,221
Total	262,432	213,280	49,152

A.4 Training Prompt and SFT Target

SFT and RL use the same deterministic prompt renderer. Each image is first enumerated as `Image 1`, `Image 2`, and so forth, after which the question is appended. If options are present, they are sorted by label. An image-valued option is rendered as a reference to its corresponding image position rather than as a raw file path. Single-choice prompts request one option letter, whereas multi-select prompts request comma-separated option letters. Every prompt ends with the shared response instruction:

> Respond with exactly one `<reasoning>...</reasoning>` block followed by exactly one `<answer>...</answer>` block. Keep the final answer concise and deterministically parseable.

For SFT, the target contains an empty reasoning block followed by the gold answer:

```
<reasoning></reasoning><answer>GOLD_ANSWER</answer>
```

AgroOmni provides gold answers but not gold reasoning trajectories; therefore, the SFT builder does not synthesize chain-of-thought supervision. For RL, the same user prompt is paired with the gold answer and answer type in the rule-based reward specification rather than with a supervised response target.

B Transformation Rules and Answer Canonicalization

B.1 Acquisition Views and Paired Transformations

The acquisition view specifies whether an observation comes from a ground camera, UAV, or satellite. VECPO never transforms one acquisition view into another. Instead, for each RL source it constructs an original branch and a within-view transformed branch. The transformation is selected deterministically from horizontal flip, vertical flip, and 180° rotation using the sample identifier and training seed. In a multi-image example, the same transformation is applied to every image or frame while their order is preserved. Image-valued options are transformed in place, so their option labels are not permuted.

Let a source example be $x = (\mathbf{I}, z, y)$, where $\mathbf{I}$ is the image sequence, $z = (q, \mathcal{O})$ contains the question and optional answer set, and y is the target. For $g \in \mathcal{G} = \{H, V, R_{180}\}$, denoting horizontal

Table 5: Target transformation and canonicalization rules for VECPO paired rollouts.

Case	Target y'	Canonical answer $C_g(\hat{y}')$
Invariant	y	$\hat{y}'$
Boolean	y	$\hat{y}'$
Direction	$D_g(y)$	$D_g^{-1}(\hat{y}')$
Grid	$G_g(y)$	$G_g^{-1}(\hat{y}')$
Spatial choice	$\pi_g(y)$	$\pi_g^{-1}(\hat{y}')$
Coordinate input	y	$\hat{y}'$
Bounding box	$B_g(y)$	$B_g^{-1}(\hat{y}')$
Image choice	y	$\hat{y}'$

flip, vertical flip, and 180° rotation, the paired input is $\mathbf{I}' = g(\mathbf{I})$ and $z' = T_g(z)$. The shared input map T_g rewrites supported spatial language, transforms point or box coordinates on the original canvas, and transforms image-valued options in place; unaffected fields remain unchanged. Only the target-side map depends on the answer form. We write D_g, G_g, and B_g for the directional, grid, and bounding-box maps, and π_g for the semantic option-label permutation. The canonicalizer C_g maps a transformed prediction $\hat{y}'$ back to the original answer frame, as summarized in Table 5.

B.2 Transformation Rules by Target Type

Invariant targets. For category, count, numeric, and other non-spatial answers, the target value is unchanged. Nevertheless, if the question restricts the answer to a spatial region, supported terms such as left/right, upper/lower, north/south, and east/west are rewritten together with the image. Boolean spatial relations are handled in the same way: jointly transforming the visual input and relational language preserves the truth value. For image-valued choices, every candidate image is transformed in place, so the correct option label remains unchanged.

Discrete spatial targets. Directional short-text answers are transformed using the same directional dictionary as the question. For a row-major 3×3 grid, let a cell be (r, c) with $r, c \in \{0, 1, 2\}$ and index $3r + c + 1$. Horizontal flip maps the cell to $(r, 2 - c)$, vertical flip to $(2 - r, c)$, and 180° rotation to $(2-r, 2-c)$. Spatial multiple-choice targets are handled through option semantics: after the visual transformation, the gold label is changed to the unique option expressing the transformed direction. A spatial-choice example is eligible only when this correspondence is complete, one-to-one, and invertible for all option labels.

Coordinate-bearing prompts and targets. VECPO recognizes two-coordinate points and four-coordinate boxes appearing in questions or text-valued options. Their coordinates are transformed using the width W and height H of the original image; normalized coordinates use the unit square. For a point $p = (x, y)$,

$$M_{\text{hflip}}(p) = (W - x, y), \tag{12}$$
$$M_{\text{vflip}}(p) = (x, H - y), \tag{13}$$
$$M_{180}(p) = (W - x, H - y). \tag{14}$$

For a bounding box $b = (x_1, y_1, x_2, y_2)$,

$$M_{\text{hflip}}(b) = (W - x_2, y_1, W - x_1, y_2), \tag{15}$$
$$M_{\text{vflip}}(b) = (x_1, H - y_2, x_2, H - y_1), \tag{16}$$
$$M_{180}(b) = (W - x_2, H - y_2, W - x_1, H - y_1). \tag{17}$$

Coordinate tuples in the question and options are rewritten in place while preserving their surrounding punctuation and numeric precision. For a bounding-box task, the gold box is transformed on the same canvas. This keeps the transformed image, prompt, options, and target geometrically aligned.

Table 6: Task-specific reward functions used during policy optimization.

Answer form	Reward definition
Single choice	$a = \mathbf{1}[\ell(\widehat{y}) = \ell(y)]$
Boolean	$a = \mathbf{1}[\mathrm{norm}(\widehat{y}) = \mathrm{norm}(y)]$
Count / numeric	$a = \max\Big(0, 1 - \frac{\lvert\widehat{n}-n\rvert}{\max(\lvert n\rvert,1)}\Big)$
Multi-select / short text	$a = \frac{1}{2}\big(\mathrm{Jac}(P,Y) + \mathrm{Rec}(P,Y)\big)$
Bounding box	$a = \mathrm{IoU}(\widehat{y}, y)$
Format auxiliary	$f = \mathbf{1}[\text{valid reasoning–answer structure}]$
Grounding auxiliary	$h = \mathrm{IoU}(\widehat{y}, y)$ for boxes, and 0 otherwise
Total base reward	$r^{\mathrm{base}} = a + 0.1f + 0.2h$

C Reward Functions

C.1 Task-Aware Base Reward

During policy optimization, every completion is evaluated by a task-accuracy term and two lightweight auxiliary terms that capture output validity and spatial grounding. VECPO uses this same task-aware base reward for the matched policy-optimization baselines; its method-specific contribution is added only through the equivariance term described below. Let $\widehat{y}$ and y denote the parsed prediction and target, and let P and Y be their normalized set representations when the answer is set-valued. Table 6 summarizes the complete base-reward definition.

Task-accuracy reward. The accuracy component a is adapted to the structure of each answer rather than forcing all tasks through string equality. For single-choice questions, $\ell(\cdot)$ extracts the option label and falls back to normalized text when a label is unavailable. Boolean answers use normalized equality directly. For count and numeric questions, the relative-error form assigns partial credit to nearby values while remaining in $[0, 1]$; the denominator $\max(|n|, 1)$ also prevents instability when the target is zero. For multi-select and short-text answers, Jaccard similarity penalizes spurious elements while recall penalizes omissions, and their mean balances the two failure modes. Bounding-box answers use IoU so that the signal changes continuously with localization quality.

Output-validity reward. The format indicator f equals one only when the completion contains exactly one reasoning block followed by exactly one answer block. It encourages a stable response interface for parsing and evaluation, but it remains auxiliary: if an answer can still be extracted from a completion that fails the full-format check, its task-accuracy reward is retained. A missing or ambiguous answer receives zero task accuracy.

Spatial-grounding reward. The grounding component h is active only for bounding-box tasks and equals the IoU between the predicted and target boxes. It supplies an additional geometric signal beyond the generic task score, whereas non-box tasks set $h = 0$. Consequently, a bounding-box completion receives a combined geometric contribution of $1.2\,\mathrm{IoU}(\widehat{y}, y)$ before the format and equivariance terms are considered.

Combined base reward. The three components are combined as

$$r_i^{\mathrm{base}} = a_i + 0.1f_i + 0.2h_i. \tag{18}$$

The task score remains the dominant signal, while the smaller auxiliary weights reward valid structure and precise grounding without overwhelming answer correctness. These shaped rewards are optimization signals and do not redefine the exact-match accuracy used by the benchmark.

C.2 Correctness-Gated Equivariance Reward

Strict correctness signal. The continuous scores above are useful for optimization, but they are too permissive to decide whether cross-branch agreement should be rewarded. VECPO therefore computes a separate binary variable u_i for each response. Single-choice answers require an accepted option label; Boolean and short-text answers use normalized equality; multi-select answers require

Table 7: SFT training configuration.

Setting	Value
Seed / epochs	42 / 1
Full-parameter training	Enabled
Per-device batch / gradient accumulation	1 / 8
Learning rate / weight decay	10^{-6} / 0.01
Warmup ratio / maximum gradient norm	0.03 / 1.0
Maximum sequence length	4,096
Image pixel bounds	65,536–524,288
Gradient checkpointing	Enabled
Fast processor	Disabled

exact set equality; numeric answers match after benchmark-aligned rounding; and a bounding box is correct when its IoU is strictly greater than 0.5. Thus, a nearby number or partially overlapping set can receive nonzero base reward without opening the equivariance gate.

Canonicalized cross-branch agreement. Correctness is evaluated against each branch's own target. After mapping every prediction back to the source answer frame, let $\mathrm{cross}(i) = \{j : b(j) \neq b(i), m_{z,b(i)} m_{z,b(j)} = 1\}$ contain responses from the other eligible transformation branches. The consistency bonus for response i is

$$c_i = \begin{cases} \dfrac{1}{|\,\mathrm{cross}(i)|} \displaystyle\sum_{j \in \mathrm{cross}(i)} u_i u_j E_c(\widetilde{y}_i, \widetilde{y}_j), & |\,\mathrm{cross}(i)| > 0, \\ 0, & \text{otherwise}, \end{cases} \tag{19}$$

where $m_{z,b}$ is the branch-eligibility mask and E_c is the answer-type-specific agreement function. Averaging over all eligible cross-branch responses reduces sensitivity to any one sampled completion. The product $u_i u_j$ requires both responses to be correct, so branches that agree on the same wrong answer receive no bonus. A transformation without an exact inverse answer map is omitted or masked from this set.

Final VECPO reward. The consistency coefficient is introduced gradually,

$$\alpha_t = \alpha_{\max} \min\left(1, \frac{t}{T_\alpha}\right), \tag{20}$$

and the final response reward is

$$r_i = r_i^{\mathrm{base}} + \alpha_t c_i. \tag{21}$$

The warmup prevents the consistency term from dominating before the policy produces sufficiently reliable answers. In our configuration, $\alpha_{\max} = 0.2$ and $T_\alpha = 77$ optimization steps, as reported in Table 8.

D Training Configuration and Procedure

D.1 SFT Training Configuration

VECPO and all matched RL baselines start from the same Qwen2.5-VL-7B SFT checkpoint. This common initialization separates the effect of the subsequent policy-optimization rule from differences in supervised data or warm-start quality. Table 7 reports the complete SFT configuration.

Optimization setup. We perform one epoch of full-parameter supervised fine-tuning with a learning rate of 10^{-6}, weight decay of 0.01, and a 0.03 warmup ratio. The per-device batch size is one and gradients are accumulated for eight steps. Gradient clipping at norm 1.0 and checkpointed activations control numerical and memory pressure while retaining full-parameter updates. This stage teaches the model the shared response format and agricultural task distribution before any reward-based optimization is introduced.

Table 8: RL training configuration for VECPO.

Setting	Value
Hardware / seed	Four A100 GPUs / 42
Source prompts / branch records	49,152 / 98,304
Sources per step / responses per branch	64 / 2
Completions per step / optimization steps	256 / 768
Actor learning rate / PPO epochs	10^{-6} / 1
Actor mini-batch size	64
Actor / rollout / reference micro-batch per GPU	2 / 4 / 4
Prompt / response / model length limit	4,096 / 512 / 8,192
Rollout temperature / top-p	1.0 / 0.99
Actor KL coefficient / entropy coefficient	10^{-4} / 0
Maximum consistency coefficient / warmup	0.2 / 77 steps
DGCA source–branch mixture	0.5 / 0.5
Controller warmup / update interval	32 / 8 steps
EMA decay / multiplier learning rate	0.95 / 0.05
Constraint tolerance / multiplier cap	0.01 / 1.0
Normalized weight range	$[0.5, 2.0]$

Multimodal input configuration. The maximum sequence length is 4,096 tokens. Images are processed within a 65,536–524,288 pixel range, which bounds visual-token cost while preserving the detail needed by ground, UAV, and satellite observations. The fast image processor is disabled in the checked-in configuration, keeping preprocessing deterministic for this warm start.

D.2 RL Training Configuration

Table 8 lists the VECPO training configuration. The configuration jointly controls the transformation-branch budget, policy update, equivariance schedule, and group controller. WGCC estimates exact-correctness anchors from training-time warm-start statistics and freezes them after the warmup; the final evaluation set is not used to construct these anchors.

Reported transformation-branch budget. The 49,152 RL sources are expanded into 98,304 branch records by pairing each original input with one transformed counterpart. At every step, 64 sources produce 128 branch prompts, and two responses are sampled for each branch, yielding 256 completions. Over 768 optimization steps, the sampler therefore covers the full source budget once while preserving all four responses from each source as one comparison group. This identity-plus-one-transform layout, with a masked identity fallback for unsafe transformations, is the reported instantiation of the general multi-branch formulation.

Generation and policy optimization. Rollouts use temperature 1.0 and top-$p = 0.99$ to provide within-group variation for relative advantage estimation. The actor is updated for one PPO epoch with learning rate 10^{-6} and mini-batch size 64. Prompt, response, and model limits are set to 4,096, 512, and 8,192 tokens, respectively. An actor-side KL coefficient of 10^{-4} regularizes the policy toward the SFT reference, while the entropy coefficient is zero. The actor, rollout, and reference micro-batches remain fixed at $2/4/4$ per GPU on four A100 GPUs.

DGCA and WGCC. The equivariance coefficient increases to 0.2 during the first 77 steps. DGCA first couples source-level and within-branch advantages equally, using source-level normalization over the four implemented responses together with branch-local comparisons. WGCC begins after a 32-step warmup and updates every eight steps. It tracks group outcomes with EMA decay 0.95, applies a 0.01 tolerance, and projects its multiplier to $[0, 1]$ using learning rate 0.05. Response weights are subsequently normalized and clipped to $[0.5, 2.0]$, completing the constrained advantage without allowing a single group to set the global advantage scale.

D.3 VECPO Training Pseudocode

Algorithm 1 summarizes the complete VECPO loop. The ordering is important: transformed responses must be canonicalized before cross-branch agreement is measured, DGCA must be formed

Algorithm 1 VECPO training procedure

Require: Policy π_θ, reference π_{ref}, transformation-branch data $\mathcal{D}$, and controller hyperparameters

for optimization step $t = 1, \ldots, T$ **do**
  Phase 1: Multi-branch rollout and reward computation
    Sample sources and construct eligible transformation branches
    Draw n_b responses from each branch using $\pi_{\theta_{\text{old}}}$
    Compute r_i^{base}, correctness u_i, and canonical prediction $\widetilde{y}_i$ for each response
    Compute c_i using equation 19 and r_i using equation 21
  Phase 2: DGCA and WGCC
    Compute source-level advantage A_i^{src} and within-branch advantage A_i^{br}
    $A_i^{\text{DG}} \leftarrow (1-\rho)A_i^{\text{src}} + \rho A_i^{\text{br}}$
    Update exact-correctness EMAs for observed groups
    if t is within the controller warmup **then**
      Update the training-derived anchors b_g
    else if the controller update is scheduled **then**
      Update projected multipliers λ_g relative to b_g
    end if
    Compute normalized group weight w_i and set $\widehat{A}_i^{\text{DGCA}} \leftarrow w_i A_i^{\text{DG}}$
  Phase 3: Policy update
    Update θ with the clipped objective using $\widehat{A}_i^{\text{DGCA}}$ and KL regularization toward π_{ref}
end for

before WGCC applies group weights, and controller statistics must be updated before the policy step.

Phase 1: multi-branch rollout and reward computation. For each sampled source, all eligible transformation branches are kept in the same optimization group and receive independently sampled responses. The reported configuration uses the identity and one transformed branch with two responses each. The task-aware base reward is evaluated against the branch-local target, while strict correctness supplies the gate used by the equivariance term. Predictions are then mapped back to the source reference frame before cross-branch agreement is computed. This ordering prevents a valid geometric transformation from being mistaken for disagreement.

Phase 2: DGCA and WGCC. The source-level advantage compares all responses derived from one source, whereas the branch advantage compares responses generated from the same transformed prompt. In the reported configuration, these sets contain four and two responses, respectively. DGCA couples them using ρ, retaining shared cross-branch comparison while preserving within-branch discrimination. In parallel, WGCC updates exact-correctness EMAs for every view–task–answer-type group. During the initial warmup, the EMA of each observed group also defines its anchor; these anchors are then frozen. At scheduled controller steps, deviations below the anchors update the projected multipliers, which are converted into batch-normalized and clipped response weights.

Phase 3: policy update. WGCC applies the normalized group weights to the dual-granularity advantages. The actor is then optimized with the clipped policy objective using $\widehat{A}_i^{\text{DGCA}}$ together with KL regularization toward the fixed SFT reference policy.

D.4 Reinforcement-Learning Baselines

We compare VECPO with five representative reinforcement-learning optimizers. All methods are initialized from the same Qwen2.5-VL-7B SFT checkpoint, while each baseline retains its own advantage estimator and policy objective.

- **PPO** (Schulman et al., 2017) is the actor–critic baseline. It learns a value model alongside the policy and uses the resulting value estimates to construct low-variance advantages. Its clipped importance-ratio objective limits excessively large policy changes, while KL regularization keeps

Table 9: AgroMind task families and audited evaluation sample counts. Task identifiers follow the benchmark output files.

Family	Task	Examples
Spatial perception	SL	2,036
	SR	1,286
	BD	1,820
Object understanding	OC	3,477
	PDD	2,300
	GSR	1,200
Scene understanding	SC	1,657
	CO	4,231
	AS	3,176
Scene reasoning	AR	1,521
	CTR	1,073
	VPR	2,364
	PL	2,341
Total		28,482

the updated policy close to the SFT reference. PPO provides the principal comparison between VECPO's critic-free design and value-based optimization.

- **GRPO** (Shao et al., 2024) removes the learned critic. It samples multiple responses for each prompt and constructs advantages by centering and scaling their rewards within the response group. The resulting group-relative signal is shared across the tokens of each response and is optimized with a clipped policy objective and reference-policy regularization. GRPO is the closest generic critic-free baseline to VECPO.
- **ReMax** (Li et al., 2023) is a REINFORCE-style critic-free method designed around trajectory-level rewards. For each prompt, it generates a stochastic response for learning and a greedily decoded response whose reward serves as a prompt-dependent baseline. Subtracting this baseline reduces policy-gradient variance without training a value model or estimating group-normalized advantages.
- **RLOO** (Ahmadian et al., 2024) constructs a leave-one-out control variate from multiple responses to the same prompt: each response is compared with the mean reward of all remaining responses. This estimator preserves a prompt-specific baseline without a learned critic or a separate greedy decoding pass. It therefore provides a direct comparison between leave-one-out variance reduction and the group-standardized advantages used by GRPO.
- **REINFORCE++** (Hu et al., 2025) is a critic-free extension of REINFORCE that incorporates stabilization techniques commonly used in PPO while omitting the value network. It uses token-level returns, masked advantage normalization, clipped policy updates, and KL regularization toward the reference policy. This baseline tests whether these general optimization refinements alone account for VECPO's behavior.

E AgroMind Benchmark and Qualitative Cases

E.1 AgroMind Benchmark

AgroMind (Li et al., 2026) is a heterogeneous agricultural benchmark containing 28,482 image–question–answer pairs and 20,850 images collected from nine public datasets and one private farmland-parcel dataset. It covers ground, unmanned-aerial-vehicle, and satellite observations and supports judgment, counting, image-choice, text-choice, and open-ended questions. Rather than reducing agricultural understanding to recognition alone, AgroMind organizes 13 tasks into a progression from spatial perception and object understanding to scene understanding and scene reasoning. Table 10 lists the complete taxonomy; the corresponding evaluation-set sizes are given in Table 9.

Table 10: The 13 AgroMind task types grouped by evaluation dimension.

Dimension	ID	Task type	Evaluation focus
Spatial perception	SL	Spatial Localization	Locating agricultural objects or regions within an image.
	SR	Spatial Relationship	Determining relative positions between objects or anomalous areas.
	BD	Boundary Detection	Recovering coordinates that enclose cultivated regions.
Object understanding	OC	Object Classification	Identifying crops, fruits, pests, and other agricultural objects.
	PDD	Pest/Disease Diagnostics	Recognizing plant health, pests, and disease conditions.
	GSR	Growth Status Recognition	Assessing the visible growth or condition of plants.
Scene understanding	SC	Scene Comparison	Comparing agricultural content across multiple images.
	CO	Counting	Counting trees, tassels, fruits, or other target objects.
	AS	Area Statistics	Estimating cultivated-area or land-cover proportions.
Scene reasoning	VPR	Visual Prompt Reasoning	Reasoning from visual prompts, marked regions, or scale information.
	AR	Anomaly Reasoning	Identifying and interpreting anomalous farmland regions.
	CTR	Climate Type Reasoning	Inferring climate or ecosystem type from scene evidence.
	PL	Planning	Answering decision-oriented agricultural planning questions.

We use AgroMind exclusively for held-out evaluation. All models receive the same prompts and deterministic decoding configuration, and accuracy is computed independently for each task. The Overall result in the main paper is the unweighted average of the 13 task accuracies, so each task contributes equally regardless of its number of examples. This task-balanced aggregation is particularly useful because the benchmark ranges from structured outputs, such as coordinates and counts, to fine-grained recognition and contextual reasoning. For the qualitative analysis below, we additionally apply a safe spatial transformation to each selected observation and compare whether the prediction remains correct under the known answer correspondence.

E.2 Qualitative Cases

Figures 6–9 compare GRPO and VECPO on four AgroMind examples before and after transformation. Each panel shows the original and transformed observation, the question, and the two models' responses. Across the four cases, GRPO answers the original observation correctly but changes to an incorrect prediction after transformation, whereas VECPO remains correct on both observations. The examples span anomaly reasoning, pest recognition, fruit recognition, and disease diagnosis, illustrating transformation consistency across different agricultural tasks.

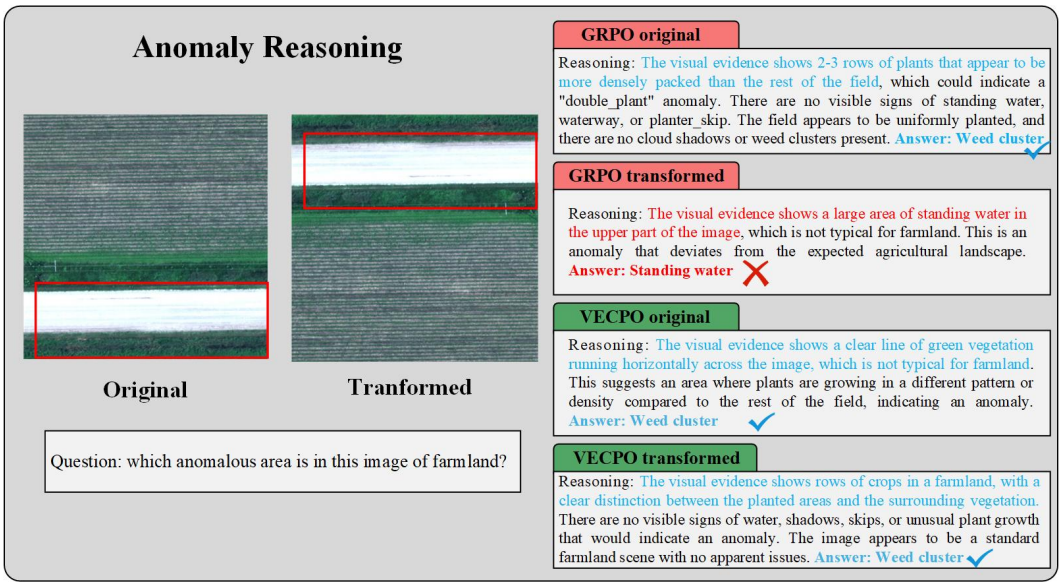


Figure 6: Anomaly-reasoning case. GRPO changes its prediction from weed cluster to standing water after the transformation, whereas VECPO retains the correct weed-cluster prediction for both observations.

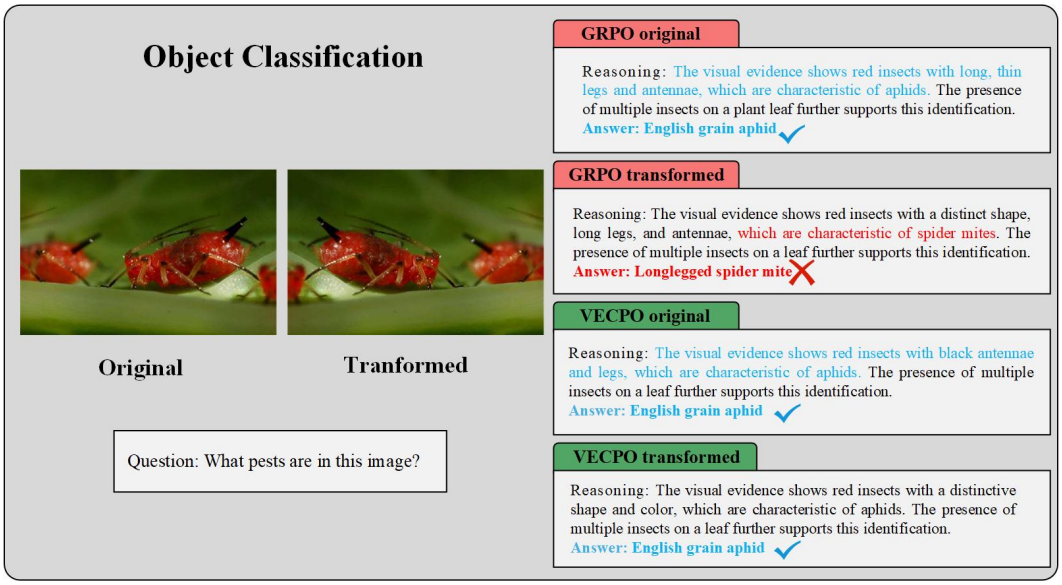


Figure 7: Pest-recognition case. GRPO changes its prediction from English grain aphid to longlegged spider mite after the transformation, while VECPO correctly identifies the pest in both observations.

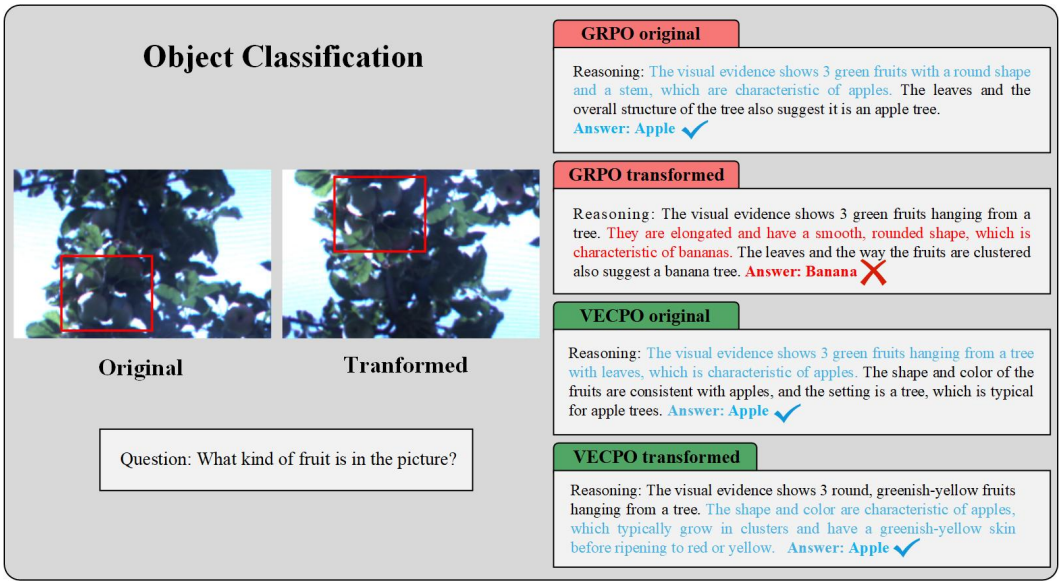


Figure 8: Fruit-recognition case. GRPO changes its answer from apple to banana after the transformation, whereas VECPO consistently gives the correct apple prediction.

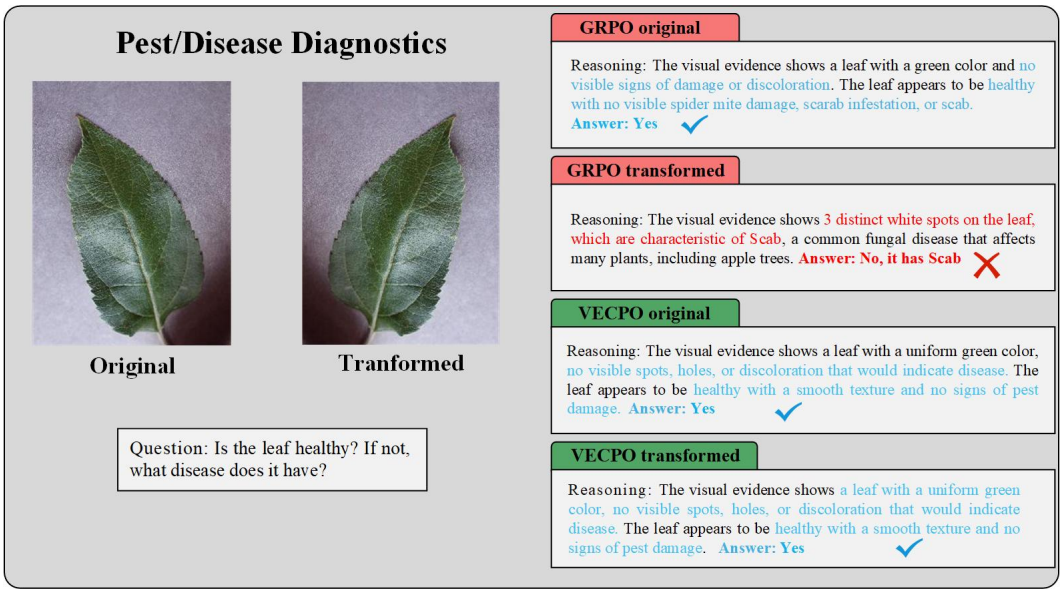


Figure 9: Disease-diagnosis case. GRPO incorrectly diagnoses scab after the transformation, while VECPO consistently recognizes the leaf as healthy.