AGRISEED: LEARNING TO TEACH AGRICULTURAL VISION–LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Generating useful supervision is a central challenge in adapting vision–language models to agricultural tasks. Although models can generate plausible question–answer pairs, answer validity alone does not indicate whether an example will help a solver learn. We introduce AGRISEED (**S**olver-**E**valuated **E**xample **D**esign), a learning-to-teach framework that optimizes question generation using feedback from temporary solver adaptation. Given agricultural images, a questioner proposes question–answer pairs, and a frozen visual judge screens their quality. For each reliable candidate, a temporary solver undergoes a single low-rank adaptation step. Its learning gain is measured through changes in reference-answer likelihood on screened, self-generated probes from other images, complemented by augmented views of the candidate image. The temporary update is reset after evaluation, and the learning gain is combined with format and answer-quality rewards to train the questioner using GRPO. The trained questioner then generates a filtered dataset for a subsequent stage of solver reinforcement learning under the fixed visual judge. Experiments on AgroCoT with Qwen2.5-VL-7B, Qwen3-VL-8B, and InternVL2-8B show consistent improvements in semantic similarity and chain-of-thought quality across all five agricultural task dimensions. A task-wise analysis shows consistent improvements in both semantic alignment and reasoning-step matching across the evaluated agricultural tasks and model backbones, supporting the effectiveness of AGRISEED for agricultural vision–language understanding.

1 INTRODUCTION

Agricultural vision–language understanding requires models to connect fine-grained visual evidence with domain knowledge, from counting plants and locating objects to interpreting disease symptoms and field conditions. Advances in multimodal pretraining and visual instruction tuning provide a promising foundation for these capabilities (Liu et al., 2023; Bai et al., 2025b), whose breadth is reflected in the five task dimensions of AgroCoT (Wen et al., 2026). Developing them, however, requires useful supervision: the many questions that an agricultural image supports need not be equally informative for a learner. Figure 1 summarizes our approach to this challenge: learning to generate supervision using feedback from the solver's learning process. The central question is *which visually grounded questions help an agricultural VLM improve?*

Synthetic instruction generation offers a scalable way to construct supervision (Wang et al., 2023; Liu et al., 2023). Agricultural systems, including AgroGPT (Awais et al., 2025), Agri-LLaVA (Wang et al., 2024), and AgriGPT-VL (Yang et al., 2025), adapt VLMs using domain knowledge and curated or synthesized instructions. Agri-R1 further combines synthetic agricultural reasoning data with GRPO (Zhang et al., 2026). These approaches improve the availability and quality of domain supervision, but a valid example is not necessarily a useful teaching example. A correct question may repeat a skill the solver already possesses, while a plausible but poorly grounded question may encourage unsupported reasoning. Selecting supervision therefore calls for evaluating both its quality and what the solver can learn from it.

Self-play makes task generation responsive to the learner. R-Zero (Huang et al., 2026) and VisPlay (He et al., 2026) adapt generated questions using solver feedback, while SPICE (Liu et al., 2025) grounds tasks and rewards questions according to the variance of current answer correctness. Such feedback helps identify challenging questions, but current difficulty does not directly measure the benefit of

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.

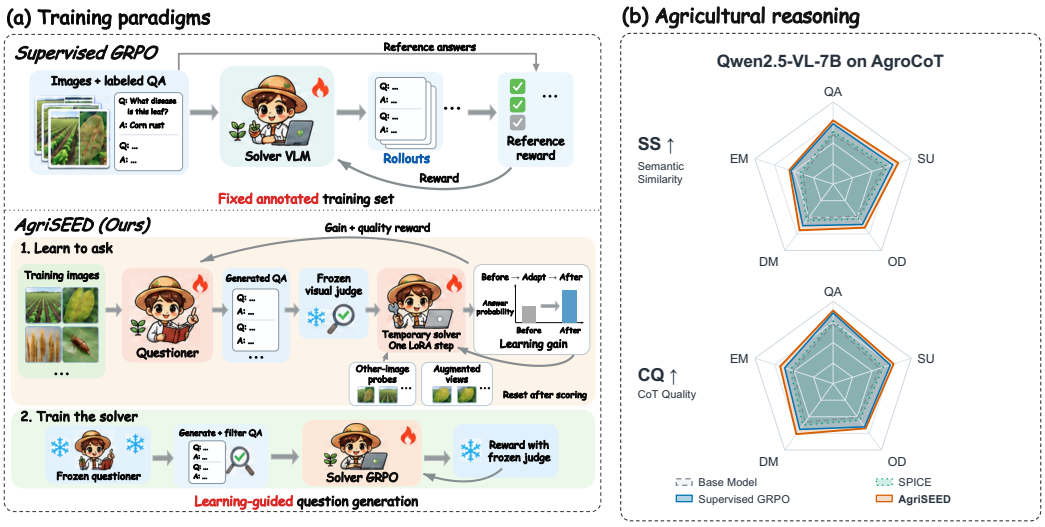

Figure 1: **Training paradigms and agricultural reasoning performance. (a)** Supervised GRPO uses a fixed annotated question–answer set. AGRISEED trains a questioner with quality feedback and learning gains measured after temporary solver adaptation, then uses its filtered questions to train the final solver. **(b)** Qwen2.5-VL-7B results on AgroCoT across quantitative analysis (QA), spatial understanding (SU), object detection (OD), disease monitoring (DM), and environmental management (EM).

learning from them. In agricultural images, uncertainty may also arise from ambiguous symptoms or insufficient visual evidence. A question that is difficult to answer can consequently provide little transferable supervision. Following the learning-to-teach principle (Fan et al., 2018), we instead assess a candidate through the solver's progress after learning from it, using this progress to guide subsequent question generation.

We introduce AGRISEED (**S**olver-**E**valuated **E**xample **D**esign), a framework for learning to generate agricultural supervision. As shown in Figure 1(a), a frozen visual judge first screens the questioner's candidate question–answer pairs. For each accepted pair, we temporarily adapt the solver with one LoRA update (Hu et al., 2022) and measure the change in reference-answer log-likelihood on screened, self-generated probes. These probes use other images when available, together with augmented views of the candidate image, to assess benefits beyond the original training input. Temporary updates are reset between candidates. The estimated gain, combined with format and answer-quality rewards, trains the questioner through GRPO (Shao et al., 2024). We then freeze the questioner, generate and filter a training set, and train the final solver in a separate GRPO stage. This procedure makes estimated teaching value an explicit objective of question generation.

We evaluate AGRISEED on AgroCoT with Qwen2.5-VL-7B (Bai et al., 2025b), Qwen3-VL-8B (Bai et al., 2025a), and InternVL2-8B (OpenGVLab Team, 2024). Averaged across the three backbones, AGRISEED improves overall semantic similarity and chain-of-thought quality over the base models by 4.75 and 4.08 points, respectively. Task-wise analysis shows gains in both metrics across all five agricultural dimensions for every backbone, indicating benefits across complementary capabilities. AGRISEED also exceeds supervised GRPO on both overall metrics for Qwen2.5-VL-7B and InternVL2-8B, with slightly lower scores on Qwen3-VL-8B. Our direct adaptation of SPICE yields only limited gains, underscoring the challenge of transferring its self-play recipe to agricultural reasoning. Together, these comparisons support learning-guided question generation as an effective approach to agricultural VLM adaptation.

2 RELATED WORK

Agricultural vision–language models. General-purpose VLMs provide a foundation for agricultural adaptation. BLIP-2 connects frozen image encoders and language models through a lightweight

bridge (Li et al., 2023), while LLaVA demonstrates the effectiveness of synthetic visual instruction data (Liu et al., 2023). In agriculture, AgroGPT constructs expert instruction data for domain adaptation (Awais et al., 2025), and Agri-LLaVA incorporates structured pest and disease knowledge (Wang et al., 2024). AgriGPT-VL combines large-scale agricultural data construction with progressive alignment and reinforcement learning (Yang et al., 2025). Agri-R1 further couples automated reasoning-data synthesis and filtering with domain-aware GRPO rewards (Zhang et al., 2026). These studies establish the value of agricultural data and specialized training objectives. Our focus is the teaching value of individual examples: the question generator is optimized using the solver's measured response to learning from its generated supervision.

Agricultural visual question answering and reasoning. Recent benchmarks assess capabilities beyond conventional crop or disease classification. AgroBench evaluates fine-grained agricultural recognition and knowledge (Shinoda et al., 2025), while AgMMU builds multimodal questions from authentic interactions between growers and agricultural experts (Gauba et al., 2025). PlantExpertVQA extends plant-image collections into expert-reviewed visual question answering (Sakib et al., 2026). AgroCoT evaluates both answers and reasoning across five agricultural task dimensions (Wen et al., 2026) and provides the evaluation setting for this work. Together, these resources motivate supervision that develops visual grounding alongside agricultural reasoning.

Synthetic supervision and multimodal self-play. Self-Instruct generates and filters instructions to improve language models (Wang et al., 2023). In the multimodal setting, SIMA uses self-generated responses and visual self-evaluation to construct preference supervision (Wang et al., 2025). Self-play additionally adapts the training tasks themselves: R-Zero trains a challenger through solver feedback (Huang et al., 2026), and VisPlay extends questioner–solver interaction to image-conditioned task generation (He et al., 2026). SPICE grounds generated tasks and rewards questions according to the variance of current answer correctness (Liu et al., 2025). Absolute Zero uses an executable environment to validate tasks and solutions (Zhao et al., 2025), while MM-Zero combines task proposal, code-based image generation, and visual problem solving (Li et al., 2026). These methods provide complementary sources of grounding and adaptive feedback. However, solver uncertainty and success-rate-based difficulty are indirect indicators of the benefit of training on a particular example. AGRISEED augments visual quality checks with the measured effect of a temporary solver update, connecting question generation to estimated learning benefit.

Learning to teach and adaptive data generation. Using student progress to guide supervision is an established research direction. Automated curriculum learning and teacher–student curriculum learning adapt task sampling using learning-progress signals (Graves et al., 2017; Matiisen et al., 2020), while Learning to Teach optimizes a teaching policy through student feedback (Fan et al., 2018). Learning to Reweight Examples and Meta-Weight-Net use clean feedback data to adjust the influence of training examples (Ren et al., 2018; Shu et al., 2019). Meta Pseudo Labels updates a pseudo-labeling teacher according to the student's subsequent performance on labeled data (Pham et al., 2021). Dataset Distillation and Generative Teaching Networks optimize synthetic training data through the learner's training process (Wang et al., 2018; Such et al., 2020). AGRISEED builds on this feedback principle for agricultural QA generation. Its feedback comes from screened, self-generated probes, including other training images, after a temporary LoRA update (Hu et al., 2022). The measured gain becomes a GRPO reward (Shao et al., 2024) for the questioner, without differentiating through the solver update. This places our contribution in the design of visual QA supervision and its teaching signal, with questioner optimization separated from final solver training.

3 METHOD

3.1 OVERVIEW

AGRISEED learns to generate agricultural supervision by evaluating what a solver learns from each candidate example. Given a pool of training images $\mathcal{I}$, we use a questioner Q_ϕ, an initial solver S_{θ_0}, and a frozen visual judge J. Figure 2 presents two sequential stages. First, we optimize the questioner using visual quality feedback and learning gains measured after temporary solver adaptation. Second, we freeze the questioner, construct a filtered synthetic dataset, and train the final solver from θ_0. The temporary adaptations serve only to evaluate teaching value; they are reset and do not initialize the final solver. Both stages obtain their QA supervision from generated references rather than dataset-provided answers.

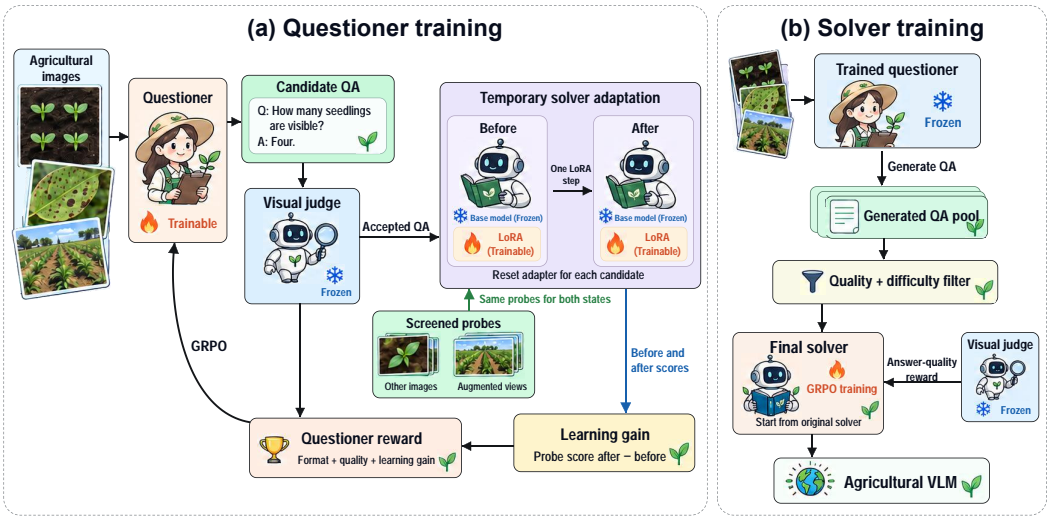

Figure 2: **Overview of** AGRISEED. **(a) Questioner training.** A trainable questioner generates agricultural QA pairs. A frozen visual judge assesses their quality. Reliable candidates induce one temporary LoRA update in a solver with a frozen base model. The same screened probes are evaluated before and after the update to estimate learning gain, which joins format and quality feedback in the questioner's GRPO reward. The adapter is reset for every candidate. Probes contain screened questions from other training images when available and augmented views of the candidate image; all probe evaluations use fixed augmentations. **(b) Solver training.** The frozen, trained questioner generates a QA pool. Reference verification and difficulty filtering produce a fixed training set, on which the final solver learns through GRPO with a frozen judge's answer-quality reward. The final solver starts from its original checkpoint.

3.2 CANDIDATE GENERATION AND SCREENING

For each image x, the questioner samples a group of structured outputs containing a question q and a reference answer $\hat{a}$. We denote a candidate by $e = (x, q, \hat{a})$. The generation prompt asks for questions supported by visible evidence and concise agricultural answers. A format score $F_e \in [0, 1]$ rewards recoverable structure and complete fields. Candidates with a parseable QA pair, a readable image, and a sufficiently long question proceed to visual assessment.

The judge evaluates $\hat{a}$ against (x, q) without an external reference. It scores correctness, completeness, visual grounding, and relevance, normalized to $s_a, s_c, s_v, s_r \in [0, 1]$. Its quality score is

$$\begin{aligned} c_J &= 0.50s_a + 0.25s_c + 0.20s_v + 0.05s_r, \\ r_J(x, q, y; r) &= \min\{s_a, c_J\}, \end{aligned} \tag{1}$$

where y is the evaluated answer and r is an optional reference. Invalid answers, severe errors, or unsafe recommendations receive zero. The correctness cap prevents other dimensions from compensating for an incorrect answer. For a generated reference, $C_e = r_J(x, q, \hat{a}; \varnothing)$. A reliable candidate must also be judged answerable, satisfy $C_e \geq 0.80$, and obtain correctness and grounding scores of at least 0.75 each. Only reliable candidates trigger temporary adaptation. Other parseable candidates retain their quality feedback and receive a neutral gain score, so the questioner can improve before consistently producing reliable QA.

3.3 MEASURING LEARNING BENEFIT

Temporary adaptation. We attach a small LoRA adapter (Hu et al., 2022) with parameters u to the frozen solver, writing the resulting model as $S_{\theta_0,u}$. For a reliable candidate with T_e answer tokens,

we compute the average answer loss and take one clipped SGD step:

$$\ell(u;e) = -\frac{1}{T_e}\sum_{t=1}^{T_e} \log S_{\theta_0,u}(\hat{a}_t \mid x, q, \hat{a}_{<t}),$$
$$u_e^+ = u_0 - \eta d_e, \qquad d_e = \mathrm{clip}_c(\nabla_u \ell(u_0;e)). \tag{2}$$

Here clip_c denotes global gradient-norm clipping. The update uses the original candidate image. Every candidate starts from the same saved adapter state u_0, which is restored after evaluation. This makes the feedback comparable without accumulating candidate-specific changes in the evaluation solver.

Screened probes. To assess benefits beyond fitting the candidate itself, we construct a probe set $\mathcal{P}_e$ from the current generation batch. We select one high-quality representative per image, check its reliability, and choose up to two reliable representatives from other images. Image-content identifiers exclude the candidate's image. Candidates from the same image share the same cross-image probes. These targets are screened, self-generated references, not human-labeled validation answers.

For any QA example p, let $\bar{\ell}(u;p)$ be its answer loss averaged over fixed, mild image augmentations. We use appearance perturbations such as small color changes, blur, and noise, intended to preserve the answer. The views remain identical before and after adaptation and across candidates from the same image. We combine cross-image and same-image evaluation as

$$\Psi_e(u) = \lambda \operatorname*{median}_{p\in\mathcal{P}_e} \bar{\ell}(u;p) + (1-\lambda)\bar{\ell}(u;e),$$
$$\Delta_e = \Psi_e(u_0) - \Psi_e(u_e^+). \tag{3}$$

The default cross-image weight is $\lambda = 0.75$. When no reliable cross-image probe is available, we omit the median term and use $\Psi_e(u) = \bar{\ell}(u;e)$. Thus, Δ_e measures a reduction in length-normalized probe loss, equivalently an increase in the corresponding aggregate reference-answer log-likelihood. The median is computed separately before and after adaptation. With one or two cross-image probes, it reduces to the single value or the mean of the two values, respectively.

3.4 Questioner Optimization and Final Solver Training

Learning the questioner. We map the signed gain to a bounded score and combine it with format and reference quality:

$$G_e = \tfrac{1}{2}[1 + \tanh(\Delta_e/\tau)],$$
$$R_Q(e) = w_f F_e + w_c C_e + w_g G_e. \tag{4}$$

We use $\tau = 0.25$, so $G_e = 0.5$ denotes zero measured gain. Eligible but unreliable candidates skip adaptation and receive $G_e = 0.5$. To prioritize usable outputs early in training, the weights depend on the batch's eligible-output rate v. With $p = \min(v/0.75, 1)$, we set $(w_f, w_c, w_g) = (1 - 0.75p, 0.30p, 0.45p)$. Here eligibility concerns parsing and input checks, not judge approval. Ineligible outputs receive only $w_f F_e$. Once $v \geq 0.75$, the weights become $(0.25, 0.30, 0.45)$.

For each group of M outputs from the same image, GRPO (Shao et al., 2024) uses group-relative advantages

$$A_i = \frac{R_i - \mu_R}{\sigma_R + \epsilon}, \qquad \mu_R = \frac{1}{M}\sum_{j=1}^{M} R_j, \tag{5}$$

where σ_R is the group's population standard deviation. We optimize the questioner with clipped policy-ratio updates and a separate KL penalty to a frozen reference policy. The measured gain is a detached scalar reward: optimization does not differentiate through the temporary solver update.

Constructing supervision and training the solver. After questioner training, we freeze $Q_{\phi^\star}$ and use it to generate a QA pool. We verify the generated references with the same visual rubric. The initial solver then samples K answers for each candidate, giving a difficulty statistic

$$d(e) = \frac{1}{K}\sum_{k=1}^{K} r_J(x, q, y_k; \hat{a}), \qquad y_k \sim S_{\theta_0}(\cdot \mid x, q). \tag{6}$$

The default filter uses $K = 10$, retains verified references with $0.30 \leq d(e) \leq 0.70$ and at least 80% valid answers, and removes duplicate QA entries. This selects examples of intermediate difficulty for final training; the reference remains the verified questioner answer. We initialize the final solver from θ_0 and train it on the resulting fixed dataset using GRPO. Its outcome reward is $R_S(y; e) = r_J(x, q, y; \hat{a})$, and its advantages follow Eq. (5). The questioner and judge remain frozen. The learning-gain reward therefore trains the generator, while the answer-quality reward trains the final solver.

3.5 A Local Interpretation of Teaching Value

The feedback follows the learning-to-teach principle (Fan et al., 2018; Pham et al., 2021): evaluate supervision through its effect on a learner. For fixed probes and augmentations, suppose Ψ_e is smooth along a sufficiently small update. A first-order expansion gives

$$\Delta_e = \eta \langle \nabla_u \Psi_e(u_0), d_e \rangle + \mathcal{O}\big(\eta^2 \|d_e\|^2\big) . \tag{7}$$

Thus, the leading term measures alignment between the candidate's update direction and the probe objective. A candidate can have high current loss yet offer little benefit if these directions are poorly aligned. Conversely, an update aligned with additional agricultural questions can reduce their aggregate loss even when current difficulty alone is uninformative. Cross-image probes assess this local transfer, while augmented views assess sensitivity to appearance changes. Quality screening is intended to reduce misleading update directions, and adapter resetting provides a common comparison point.

The interpretation applies to the raw gain before its monotone normalization. The smoothness condition holds locally when the active probe losses are smooth and the examples defining a median remain unchanged; with at most two cross-image probes, the median is simply a mean. This explains how the reward can favor locally useful supervision.

4 Experiments

4.1 Experimental Setup

Benchmark and adaptation data. We study agricultural vision–language reasoning on AgroCoT (Wen et al., 2026), which contains 4,759 QA examples spanning 15 tasks in five dimensions: quantitative analysis (QA), spatial understanding (SU), object detection (OD), disease monitoring (DM), and environmental management (EM). We select 10% of the data within each task to form the adaptation pool. For AgriSEED, only the corresponding images are used: the original questions, answers, and reference CoTs are excluded from adaptation. We evaluate on all 4,759 AgroCoT examples, including the selected adaptation subset. This is a transductive image-adaptation setting: some evaluation images are visible during adaptation, although their original QA and CoT labels are withheld from AgriSEED. Sampling details appear in Appendix A.

Backbones and baselines. We evaluate Qwen2.5-VL-7B (Bai et al., 2025b), Qwen3-VL-8B (Bai et al., 2025a), and InternVL2-8B (OpenGVLab Team, 2024). The comparison includes the base-model scores from AgroCoT, supervised GRPO, and an adaptation of SPICE (Liu et al., 2025). Supervised GRPO uses the same selected data with its original labels, so its full-benchmark evaluation includes labeled training examples. In AgriSEED, the questioner generates QA targets and the frozen judge assesses their reliability. Qwen3.5-27B is used as our judge model for all the backbones. The comparison measures adaptation under these different supervision conditions, rather than performance on an image-disjoint held-out test set. We report all three iterations of AgriSEED. Both questioner and solver carry their learned checkpoints into the next iteration; temporary solver adaptations are still reset independently for each candidate within a round.

Implementation. We implement AgriSEED with verl (Sheng et al., 2025) and use GRPO (Shao et al., 2024) for both roles. Training uses eight NVIDIA A100 80GB GPUs, with two additional GPUs hosting the feedback solver during questioner optimization. Both roles use a global batch size of 64, learning rate 10^{-6}, and sampling temperature 1.0. The questioner and solver rollout group sizes are 4 and 5, respectively. Additional configuration details appear in Appendix A. Adapted-model results are means and standard deviations over three independent runs; Base values and their

Table 1: **Full-benchmark AgroCoT evaluation across three backbones.**

Method	QA		SU		OD		DM		EM		Overall	
	SS	CQ	SS	CQ	SS	CQ	SS	CQ	SS	CQ	SS	CQ
Qwen2.5-VL-7B												
Base Model	78.10	38.13	79.55	35.01	74.64	32.78	75.26	34.53	71.49	32.75	76.61	35.05
Supervised GRPO	82.05	41.50	82.89	38.32	78.23	36.30	78.79	37.25	76.13	35.48	80.42	38.18
SPICE (Liu et al., 2025)	79.18	39.07	80.87	36.09	75.81	33.77	76.51	35.65	72.72	33.72	77.82	36.07
AGRISEED	**83.28**	**42.24**	**85.02**	**39.37**	**79.69**	**36.86**	**80.87**	**39.04**	**76.78**	**36.94**	**81.93**	**39.30**
Qwen3-VL-8B												
Base Model	77.77	42.72	68.83	38.65	80.14	41.31	75.53	43.23	80.46	46.39	74.83	41.17
Supervised GRPO	**82.56**	**46.36**	73.23	42.46	**84.73**	**45.57**	**80.41**	**47.38**	**86.15**	**51.08**	**79.70**	**45.28**
SPICE (Liu et al., 2025)	78.69	43.90	69.98	39.96	81.15	42.53	76.60	44.59	81.51	47.67	75.87	42.44
AGRISEED	82.29	46.34	**73.74**	**42.60**	84.52	45.02	80.29	47.31	85.44	50.18	79.54	45.00
InternVL2-8B												
Base Model	80.60	39.43	81.95	39.01	78.66	38.42	75.55	39.46	75.85	39.25	79.11	39.12
Supervised GRPO	83.57	**43.46**	85.29	42.93	82.22	41.98	78.31	42.84	79.47	42.96	82.36	42.84
SPICE	81.46	40.31	82.99	40.02	79.61	39.35	76.63	40.53	76.82	40.16	80.09	40.08
AGRISEED	**84.65**	43.44	**86.37**	**43.34**	**82.82**	**42.50**	**79.92**	**43.73**	**80.00**	**43.31**	**83.34**	**43.27**

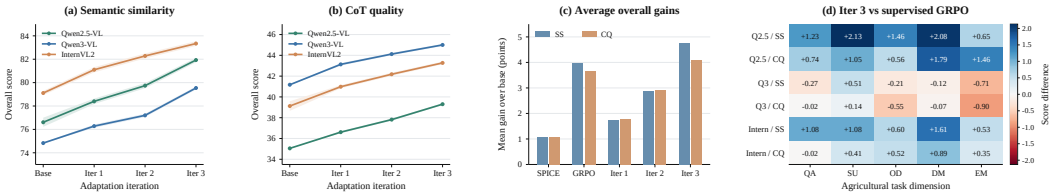

Figure 3: **Benefits of iterative adaptation.** Both SS and CQ improve across all three evaluated iterations for every backbone.

reported dispersion are reproduced from AgroCoT. For supervised GRPO and SPICE, the default hyperparameters provided in their open-source implementations are used.

Evaluation metrics. We adopt AgroCoT's two complementary CoT metrics. Semantic Similarity (SS) compares the generated CoT c with its reference $c^\star$ using cosine similarity between their 384-dimensional `all-MiniLM-L6-v2` sentence embeddings (Reimers & Gurevych, 2019). CoT Quality (CQ), inspired by MME-CoT (Jiang et al., 2025), measures key-step matching through an F1 score. If S_{out} and S_{ref} denote the numbers of generated and reference steps and S_{match} the matched steps, the displayed scores are

$$\begin{aligned} \text{SS} &= 100\,\frac{f(c)^\top f(c^\star)}{\|f(c)\|\,\|f(c^\star)\|}, \\ \text{CQ} &= 100\,\frac{2S_{\text{match}}}{S_{\text{out}}+S_{\text{ref}}}. \end{aligned} \tag{8}$$

SS captures semantic agreement between reasoning texts, while CQ evaluates the coverage and precision of their reasoning steps. We report both per-dimension and overall scores, following AgroCoT's reporting protocol. In Table 1, bold denotes the highest mean within each backbone and column, and shading marks the final iteration of AGRISEED. Full standard deviations are provided in Appendix C.

4.2 MAIN RESULTS

Improvements over base models. Table 1 shows substantial improvements over the published base-model results under our full-benchmark adaptation protocol. After three iterations, AGRISEED achieves higher SS and CQ across all five agricultural dimensions for each backbone, covering all

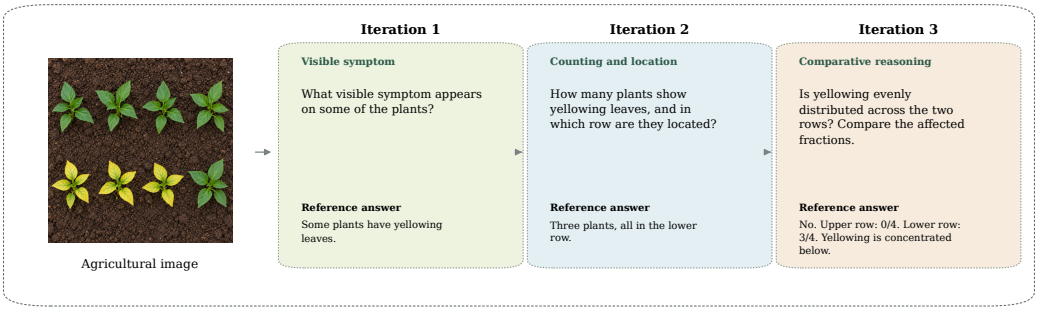

Figure 4: **Progression of agricultural questions.** This training example shows a progression from symptom recognition to counting, localization, and comparative reasoning.

30 task-level comparisons. Averaged across the three backbones, the overall improvements are 4.75 SS and 4.08 CQ points. The gains extend across quantitative, spatial, and domain-specific reasoning tasks, indicating that the benefits of generated supervision are not confined to a single agricultural capability.

Comparison with SPICE. AGRISEED also substantially outperforms the adapted SPICE self-play baseline. Its final iteration achieves average overall margins of 3.68 SS and 2.99 CQ points, with higher scores in every task–metric combination across all three backbones. Even the first iteration exceeds SPICE throughout the table, averaging 0.67 SS and 0.71 CQ points higher overall. These results establish stronger agricultural reasoning performance than the direct adaptation of SPICE in the evaluated setting; the component analysis below examines the role of learning-gain feedback.

Comparison with supervised GRPO. Although AGRISEED uses no original QA or CoT labels during adaptation, it remains competitive with supervised GRPO, which receives these labels for the same selected data. The final iteration of AGRISEED leads in 21 of the 30 task-level comparisons, improving overall SS/CQ by 1.51/1.12 points on Qwen2.5-VL-7B and 0.98/0.43 points on InternVL2-8B. Supervised GRPO retains a small overall advantage of 0.16/0.28 points on Qwen3-VL-8B, while AGRISEED performs better on both spatial understanding metrics. Thus, self-generated supervision can support strong agricultural reasoning without access to the benchmark's original QA and CoT annotations.

Effect of iterative adaptation. Figure 3 shows continued improvement as both the questioner and solver carry their learned checkpoints into successive rounds. All 30 task-level scores and six overall scores increase from Iteration 1 to 2 and again from Iteration 2 to 3. Average overall SS/CQ progresses from 78.59/40.24 to 79.74/41.37 and finally 81.60/42.52, showing that both reasoning metrics benefit throughout the three evaluated iterations. The shading in panels (a,b) represents standard deviations for our runs and the reported dispersion for the published Base entries. Panels (c,d) summarize differences from Base and supervised GRPO, respectively. This trajectory reflects the complete iterative procedure; the frozen-questioner ablation below separately examines the contribution of questioner updates.

Progression of agricultural questions. Figure 4 shows how the same image can support questions that require increasingly connected use of visual evidence. The first question identifies yellowing leaves; the second counts the affected plants and locates their row; the third compares affected fractions across rows. Each answer is grounded in the visible scene, while the questions move from recognition to spatial and quantitative reasoning. This training example supports the intended progression of supervision, rather than recorded outputs of the three trained questioner checkpoints.

4.3 ABLATION STUDY

Table 2 compares the full three-iteration system with six component variants on Qwen2.5-VL-7B. We use short variant names in the table and describe the corresponding changes below.

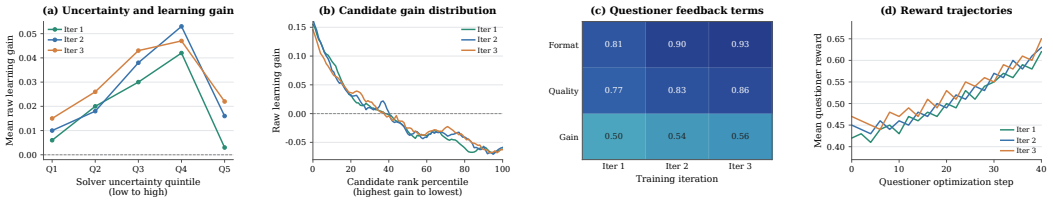

Figure 5: **Learning-gain diagnosis.** These panels organize complementary views of candidate utility, reward composition, and questioner training dynamics.

Questioner learning and reward design. *Frozen questioner* keeps the initial questioner fixed while the solver continues training; it reduces SS/CQ by 1.66/1.18 points, supporting the value of improving the generator along with the learner. The two reward variants suppress either the learning-gain or the quality contribution while preserving format feedback and the reliability gate. Removing learning-gain feedback produces the largest decline among the tested variants: 2.22 SS and 1.54 CQ points. Removing the quality reward also lowers SS/CQ by 1.07/1.16 points. Together, these results support combining feedback about reference quality with feedback about what the solver learns from an example. Finally, replacing adaptive weights with fixed reward weights yields smaller decreases of 0.21/0.13 points.

Table 2: **Component contributions.** Removing learning-gain feedback causes the largest decline in SS and CQ.

Variant	SS ↑	CQ ↑
Frozen questioner	80.27	38.12
No learning-gain reward	79.71	37.76
No quality reward	80.86	38.14
Fixed reward weights	81.72	39.17
No cross-image probes	81.02	37.98
No same-image term	81.61	39.08
AGRISEED (full)	**81.93**	**39.30**

Where learning gain is measured. The probe variants remove either the cross-image probes or the same-image contribution from the gain estimate. Without cross-image probes, the estimate relies on augmented views of the candidate image, and SS/CQ falls by 0.91/1.32 points. Removing the same-image term has a smaller effect, reducing the scores by 0.32/0.22 points. The larger CQ decrease without cross-image probes suggests that evaluating learning benefit beyond the candidate image is particularly useful for reasoning quality. The same-image term provides a smaller complementary benefit in the reported means.

Learning-gain diagnostics. Figure 5 places four complementary views of the training process alongside the component ablations. Panel (a) is designed to examine whether answer uncertainty predicts subsequent learning gain, while panel (b) shows how gains can be ranked to inspect variation in candidate utility. Panels (c,d) organize the reward components across rounds and the questioner's reward trajectories within each round.

5 CONCLUSION

We introduced AGRISEED, a framework that learns to generate useful agricultural supervision by measuring the learning gains induced by temporary solver adaptation. Combining visual quality assessment with learning-gain feedback connects question generation to the solver's learning needs, while iterative questioner and solver training supports continued improvement without using the benchmark's original QA or CoT annotations. Experiments on AgroCoT under the full-benchmark adaptation protocol demonstrate consistent improvements across three backbones and five agricultural dimensions. AGRISEED achieves average overall gains of 4.75 SS and 4.08 CQ points over the published base-model results, outperforms the adapted SPICE baseline, and remains competitive with supervised GRPO. Ablations highlight the contributions of learning-gain feedback, questioner optimization, and cross-image probes.

AI USE STATEMENT

Generative AI tools (Claude and Codex) were used as discussion and editing aids to brainstorm alternative explanations for observed experimental results, provide feedback on experimental controls and presentation, and improve the paper grammar. They were not used to implement the method, execute experiments, or generate the reported results.

ETHICS STATEMENT

This work studies agricultural vision–language learning using AgroCoT and publicly released model checkpoints, without collecting new human-subject data or annotations. The method updates model parameters using generated supervision, which may inherit factual errors and biases from the underlying models. Benchmark improvements do not establish the reliability of generated agricultural advice in practice. Outputs concerning crop diseases or farm-management decisions require validation by domain experts before operational use.

REPRODUCIBILITY STATEMENT

Section 3 describes the learning-gain reward, temporary solver adaptation, and questioner–solver training procedure. Section 4.1 and Appendix A describe the training settings, data visibility, evaluation metrics, and baseline protocols. Evaluation includes the images selected for adaptation, while their original questions, answers, and reference reasoning are withheld from AGRISEED. Appendix C reports adapted-model means and standard deviations across three independent runs; published base-model results are identified separately. Appendix D provides task-specific prompt templates, and Appendix B gives the assumptions and proof of the local learning-gain bound.

REFERENCES

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 *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 5687–5696, 2025.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, Mei Li, Kaixin Li, Zicheng Lin, Junyang Lin, Xuejing Liu, Jiawei Liu, Chenglong Liu, Yang Liu, Dayiheng Liu, Shixuan Liu, Dunjie Lu, Ruilin Luo, Chenxu Lv, Rui Men, Lingchen Meng, Xuancheng Ren, Xingzhang Ren, Sibo Song, Yuchong Sun, Jun Tang, Jianhong Tu, Jianqiang Wan, Peng Wang, Pengfei Wang, Qiuyue Wang, Yuxuan Wang, Tianbao Xie, Yiheng Xu, Haiyang Xu, Jin Xu, Zhibo Yang, Mingkun Yang, Jianxin Yang, An Yang, Bowen Yu, Fei Zhang, Hang Zhang, Xi Zhang, Bo Zheng, Humen Zhong, Jingren Zhou, Fan Zhou, Jing Zhou, Yuanzhi Zhu, and Ke Zhu. Qwen3-VL technical report, 2025a.

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, 2025b.

Yang Fan, Fei Tian, Tao Qin, Xiang-Yang Li, and Tie-Yan Liu. Learning to teach. In *International Conference on Learning Representations*, 2018.

Aruna Gauba, Irene Pi, Yunze Man, Ziqi Pang, Vikram S. Adve, and Yu-Xiong Wang. AgMMU: A comprehensive agricultural multimodal understanding benchmark. In *Advances in Neural Information Processing Systems*, volume 38, 2025. doi: 10.52202/085713-2540.

Alex Graves, Marc G. Bellemare, Jacob Menick, Rémi Munos, and Koray Kavukcuoglu. Automated curriculum learning for neural networks. In *Proceedings of the 34th International Conference on*

Machine Learning, volume 70 of *Proceedings of Machine Learning Research*, pp. 1311–1320. PMLR, 2017.

Yicheng He, Chengsong Huang, Zongxia Li, Jiaxin Huang, and Yonghui Yang. VisPlay: Self-evolving vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 26274–26284, 2026.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022.

Chengsong Huang, Wenhao Yu, Xiaoyang Wang, Hongming Zhang, Zongxia Li, Ruosen Li, Jiaxin Huang, Haitao Mi, and Dong Yu. R-Zero: Self-evolving reasoning LLM from zero data. In *The Fourteenth International Conference on Learning Representations*, 2026.

Dongzhi Jiang, Renrui Zhang, Ziyu Guo, Yanwei Li, Yu Qi, Xinyan Chen, Liuhui Wang, Jianhan Jin, Claire Guo, Shen Yan, Bo Zhang, Chaoyou Fu, Peng Gao, and Hongsheng Li. MME-CoT: Benchmarking chain-of-thought in large multimodal models for reasoning quality, robustness, and efficiency. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 27793–27830. PMLR, 2025.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 19730–19742. PMLR, 2023.

Zongxia Li, Hongyang Du, Chengsong Huang, Xiyang Wu, Lantao Yu, Yicheng He, Jing Xie, Xiaomin Wu, Zhichao Liu, Jiarui Zhang, and Fuxiao Liu. MM-Zero: Self-evolving multi-model vision language models from zero data, 2026.

Bo Liu, Chuanyang Jin, Seungone Kim, Weizhe Yuan, Wenting Zhao, Ilia Kulikov, Xian Li, Sainbayar Sukhbaatar, Jack Lanchantin, and Jason Weston. SPICE: Self-play in corpus environments improves reasoning, 2025.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *Advances in Neural Information Processing Systems*, 2023.

Tambet Matiisen, Avital Oliver, Taco Cohen, and John Schulman. Teacher-student curriculum learning. *IEEE Transactions on Neural Networks and Learning Systems*, 31(9):3732–3740, 2020. doi: 10.1109/TNNLS.2019.2934906.

OpenGVLab Team. InternVL2: Better than the best—expanding performance boundaries of open-source multimodal models with the progressive scaling strategy. Official model release, July 2024.

Hieu Pham, Zihang Dai, Qizhe Xie, and Quoc V. Le. Meta pseudo labels. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11557–11568, 2021.

Nils Reimers and Iryna Gurevych. Sentence-BERT: Sentence embeddings using siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 3982–3992. Association for Computational Linguistics, 2019. doi: 10.18653/v1/D19-1410.

Mengye Ren, Wenyuan Zeng, Bin Yang, and Raquel Urtasun. Learning to reweight examples for robust deep learning. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 4334–4343. PMLR, 2018.

Syed Nazmus Sakib, Nafiul Haque, Mohammad Zabed Hossain, and Shifat E. Arman. PlantExpertVQA: A visual question answering dataset for benchmarking vision-language models in plant science, 2026. Version 3; originally released as PlantVillageVQA in 2025.

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.

Guangming Sheng, Chi Zhang, Zilingfeng Ye, Xibin Wu, Wang Zhang, Ru Zhang, Yanghua Peng, Haibin Lin, and Chuan Wu. HybridFlow: A flexible and efficient RLHF framework. In *Proceedings of the Twentieth European Conference on Computer Systems*, EuroSys '25. Association for Computing Machinery, 2025. doi: 10.1145/3689031.3696075.

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

Jun Shu, Qi Xie, Lixuan Yi, Qian Zhao, Sanping Zhou, Zongben Xu, and Deyu Meng. Meta-Weight-Net: Learning an explicit mapping for sample weighting. In *Advances in Neural Information Processing Systems*, volume 32, pp. 1919–1930, 2019.

Felipe Petroski Such, Aditya Rawal, Joel Lehman, Kenneth Stanley, and Jeffrey Clune. Generative teaching networks: Accelerating neural architecture search by learning to generate synthetic training data. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 9206–9216. PMLR, 2020.

Liqiong Wang, Teng Jin, Jinyu Yang, Aleš Leonardis, Fangyi Wang, and Feng Zheng. Agri-LLaVA: Knowledge-infused large multimodal assistant on agricultural pests and diseases, 2024.

Tongzhou Wang, Jun-Yan Zhu, Antonio Torralba, and Alexei A. Efros. Dataset distillation. *arXiv preprint arXiv:1811.10959*, 2018.

Xiyao Wang, Jiuhai Chen, Zhaoyang Wang, Yuhang Zhou, Yiyang Zhou, Huaxiu Yao, Tianyi Zhou, Tom Goldstein, Parminder Bhatia, Taha Kass-Hout, Furong Huang, and Cao Xiao. Enhancing visual-language modality alignment in large vision language models via self-improvement. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pp. 268–282. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.findings-naacl.15.

Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A. Smith, Daniel Khashabi, and Hannaneh Hajishirzi. Self-Instruct: Aligning language models with self-generated instructions. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 13484–13508, 2023. doi: 10.18653/v1/2023.acl-long.754.

Yibin Wen, Qingmei Li, Zi Ye, Jiarui Zhang, Xiaoya Fan, Zurong Mai, Jing Wu, Shuohong Lou, Yuhang Chen, Henglian Huang, Yang Zhang, Defeng Gu, Lingyuan Zhao, Yutong Lu, Haohuan Fu, Jianxi Huang, and Juepeng Zheng. AgroCoT: A chain-of-thought benchmark for evaluating reasoning in vision-language models for agriculture, 2026. Version 3, revised May 17, 2026; first posted in 2025.

Bo Yang, Yunkui Chen, Lanfei Feng, Yu Zhang, Xiao Xu, Jianyu Zhang, Nueraili Aierken, Runhe Huang, Hongjian Lin, Yibin Ying, and Shijian Li. AgriGPT-VL: Agricultural vision-language understanding suite, 2025.

Wentao Zhang, Mingkun Xu, Qi Zhang, Shangyang Li, Derek F. Wong, Lifei Wang, Yanchao Yang, Lina Lu, and Tao Fang. Agri-R1: Agricultural reasoning for disease diagnosis via automated-synthesis and reinforcement learning, 2026.

Andrew Zhao, Yiran Wu, Tong Wu, Quentin Xu, Yang Yue, Matthieu Lin, Shenzhi Wang, Qingyun Wu, Zilong Zheng, and Gao Huang. Absolute zero: Reinforced self-play reasoning with zero data. In *Advances in Neural Information Processing Systems*, volume 38, 2025. doi: 10.52202/085713-3534.

Table 3: Public AgroCoT annotation inventory. Distinct image-path counts overlap across rows; their union is reported in the final row.

Dimension	QA records	Image paths	10% target
QA	1,189	1,146	238
SU	1,252	1,369	250
OD	945	995	189
DM	921	1,018	184
EM	452	643	90
Total / union	4,759	4,628	951

Table 4: Training settings. Q and S denote questioner and solver training.

Setting	Q	S
Training GPUs (A100 80GB)	8	8
Additional feedback GPUs	2	–
Global batch size	64	64
Learning rate	10^{-6}	10^{-6}
Sampling temperature	1.0	1.0
GRPO group size	4	5
Maximum response tokens	2,048	4,096
KL coefficient	0.01	0.01
FSDP parameter offloading	On	On
Training rounds	3	3
Independent runs	3	3

A Experimental Protocol and Reproducibility Details

Data visibility and evaluation coverage. All adapted models are evaluated on the complete set of 4,759 AgroCoT examples, including the subset selected for adaptation. Selection uses the benchmark's task categories, with a target sampling fraction of 10% within each task. In AgriSEED, the selected examples provide images only; their original questions, answers, and reference reasoning are not used to train the questioner or solver. Generated probes also come from the adaptation image pool. Supervised GRPO additionally receives the original labels for the same selected data. Its evaluation therefore includes labeled training examples, whereas AgriSEED has image exposure without those labels. These results describe full-benchmark adaptation under the stated supervision conditions. They do not measure generalization to an image-disjoint test set.

Sample counts. The public annotation file contains 4,759 QA records referencing 4,628 distinct image paths. Image paths and QA records are different counting units: some records contain multiple images, and images can be reused across questions and task categories. Table 3 gives an inventory of the public annotations. Its 10% column is an arithmetic sampling target, not the recorded adaptation manifest. Rounding 10% separately within the five dimensions gives 426 QA records. The 15 labeled subtasks cover 2,212 records; 112 records lack a subtask identifier. Rounding separately within the 15 subtasks gives 908 records; additionally sampling the unlabeled-subtask records within each dimension gives a total target of 426.

Training configuration. Table 4 records the confirmed settings. Both roles continue from their preceding-round checkpoints. Within a round, each candidate is evaluated from the same frozen solver checkpoint and reset LoRA initialization. The visual judge is fixed throughout the three rounds. The ten-answer sampling used to construct the final dataset estimates answer quality and difficulty; it does not supply majority-vote labels. The questioner supplies the candidate reference answer.

Evaluation and source attribution. Base means and dispersion are reproduced from AgroCoT; they are not reruns in our implementation. The other rows report the supplied mean and standard deviation across three independent training runs. SS compares generated and reference reasoning using `all-MiniLM-L6-v2`; CQ uses the reference-step matching F1 score. Both are displayed on the 100-point scale. Overall values are retained as reported, following AgroCoT's reporting convention, without replacing them by the unweighted mean of the five displayed dimensions.

B Local Bound on Learning Gain

Fix a candidate, its probes, their augmentations, and the aggregation weights. Let d_e be the clipped update direction in Eq. (2), and let Ψ_e be the probe loss in Eq. (3): views are averaged, and only cross-image probes are aggregated by a median. Assume Ψ_e is differentiable with an L-Lipschitz gradient on a neighborhood containing the update segment from u_0 to $u_0 - \eta d_e$. For larger probe sets, sufficient conditions are smooth active losses with bounded curvature and unchanged median-defining examples in that neighborhood. With at most two cross-image probes, the median equals a mean and introduces no additional order condition. Then the raw gain satisfies

$$\Delta_e \geq \eta \langle \nabla \Psi_e(u_0), d_e \rangle - \frac{L\eta^2}{2} \|d_e\|^2. \tag{9}$$

Proof. The descent lemma gives

$$\Psi_e(u_0 - \eta d_e) \leq \Psi_e(u_0) - \eta \langle \nabla \Psi_e(u_0), d_e \rangle + \frac{L\eta^2}{2} \|d_e\|^2. \tag{10}$$

Subtracting from $\Psi_e(u_0)$ proves the bound. For positive alignment, a sufficiently small step makes the lower bound positive. Gradient clipping changes the evaluated direction's magnitude; the proof treats that direction as fixed and does not differentiate through clipping or adaptation.

This is a local statement about the implemented aggregate loss before reward normalization. Median branch changes can invalidate the smoothness assumption. Because the probe targets are self-generated, the bound guarantees neither improvement on every probe nor higher ground-truth accuracy or gains after final solver training.

C Complete Results with Standard Deviations

Adapted-model entries report mean (standard deviation) over three independent runs. Base means and reported dispersion are reproduced from AgroCoT. All evaluation uses the complete benchmark, including the selected adaptation subset. The original precision is retained in the accompanying data files. Displayed entries are rounded to two decimal places. Overall scores are preserved as reported.

Table 5: Complete results for Qwen2.5-VL-7B. SS / CQ pairs are reported for each agricultural dimension.

Method	QA	SU	OD	DM	EM	Overall
	78.10 (0.91)	79.55 (0.62)	74.64 (0.14)	75.26 (0.67)	71.49 (0.94)	76.61 (0.39)
Base Model†	38.13 (0.01)	35.01 (0.07)	32.78 (0.12)	34.53 (0.31)	32.75 (0.17)	35.05 (0.02)
	82.05 (0.25)	82.89 (0.31)	78.23 (0.19)	78.79 (0.28)	76.13 (0.34)	80.42 (0.22)
Supervised GRPO	41.50 (0.10)	38.32 (0.14)	36.30 (0.11)	37.25 (0.16)	35.48 (0.13)	38.18 (0.11)
	79.18 (0.28)	80.87 (0.35)	75.81 (0.21)	76.51 (0.30)	72.72 (0.38)	77.82 (0.24)
SPICE	39.07 (0.09)	36.09 (0.11)	33.77 (0.10)	35.65 (0.15)	33.72 (0.12)	36.07 (0.10)
	79.72 (0.26)	81.43 (0.33)	76.35 (0.20)	77.21 (0.29)	73.28 (0.36)	78.40 (0.23)
AgriSEED (Iter 1)	39.55 (0.09)	36.62 (0.12)	34.27 (0.10)	36.25 (0.15)	34.31 (0.13)	36.61 (0.10)
	81.01 (0.24)	82.99 (0.32)	77.71 (0.19)	78.54 (0.27)	74.44 (0.35)	79.74 (0.22)
AgriSEED (Iter 2)	40.74 (0.09)	37.95 (0.12)	35.50 (0.10)	37.41 (0.15)	35.42 (0.13)	37.82 (0.10)
	83.28 (0.23)	**85.02 (0.29)**	**79.69 (0.17)**	**80.87 (0.25)**	**76.78 (0.31)**	**81.93 (0.19)**
AgriSEED (Iter 3)	**42.24 (0.08)**	**39.37 (0.12)**	**36.86 (0.09)**	**39.04 (0.15)**	**36.94 (0.11)**	**39.30 (0.09)**

D Prompt Design for Agricultural Tasks

Prompt organization. The prompt specification for AgriSEED separates task content from the learning-gain objective. A shared questioner instruction defines the output contract, and a task-specific block specifies the visual evidence and question content for one of the five agricultural dimensions. The solver and visual judge use shared instructions across tasks. The following templates are

Table 6: Complete results for Qwen3-VL-8B. SS / CQ pairs are reported for each agricultural dimension.

Method	QA	SU	OD	DM	EM	Overall
	77.77 (0.17)	68.83 (0.13)	80.14 (0.42)	75.53 (0.02)	80.46 (0.04)	74.83 (0.04)
Base Model†	42.72 (0.17)	38.65 (0.05)	41.31 (0.09)	43.23 (0.23)	46.39 (0.43)	41.17 (0.04)
	82.56 (0.15)	73.23 (0.21)	**84.73 (0.24)**	**80.41 (0.17)**	**86.15 (0.12)**	**79.70 (0.14)**
Supervised GRPO	**46.36 (0.12)**	42.46 (0.10)	**45.57 (0.14)**	**47.38 (0.18)**	**51.08 (0.20)**	**45.28 (0.12)**
	78.69 (0.16)	69.98 (0.18)	81.15 (0.23)	76.60 (0.14)	81.51 (0.11)	75.87 (0.13)
SPICE	43.90 (0.13)	39.96 (0.10)	42.53 (0.12)	44.59 (0.16)	47.67 (0.19)	42.44 (0.11)
	79.08 (0.15)	70.39 (0.19)	81.56 (0.22)	77.06 (0.16)	81.89 (0.12)	76.28 (0.14)
AGRISEED (Iter 1)	44.51 (0.13)	40.70 (0.10)	43.19 (0.12)	45.37 (0.17)	48.33 (0.19)	43.13 (0.12)
	79.95 (0.15)	71.34 (0.19)	82.47 (0.23)	77.99 (0.16)	82.83 (0.12)	77.20 (0.14)
AGRISEED (Iter 2)	45.49 (0.12)	41.79 (0.10)	44.20 (0.13)	46.25 (0.17)	49.32 (0.19)	44.12 (0.11)
	82.29 (0.14)	**73.74 (0.19)**	84.52 (0.21)	80.29 (0.15)	85.44 (0.10)	79.54 (0.13)
AGRISEED (Iter 3)	46.34 (0.12)	**42.60 (0.09)**	45.02 (0.13)	47.31 (0.16)	50.18 (0.18)	45.00 (0.10)

Table 7: Complete results for InternVL2-8B. SS / CQ pairs are reported for each agricultural dimension.

Method	QA	SU	OD	DM	EM	Overall
	80.60 (0.37)	81.95 (0.96)	78.66 (0.75)	75.55 (0.38)	75.85 (0.12)	79.11 (0.18)
Base Model†	39.43 (0.67)	39.01 (0.04)	38.42 (0.81)	39.46 (0.18)	39.25 (0.26)	39.12 (0.46)
	83.57 (0.23)	85.29 (0.29)	82.22 (0.26)	78.31 (0.22)	79.47 (0.20)	82.36 (0.21)
Supervised GRPO	**43.46 (0.19)**	42.93 (0.16)	41.98 (0.21)	42.84 (0.17)	42.96 (0.18)	42.84 (0.15)
	81.46 (0.25)	82.99 (0.32)	79.61 (0.27)	76.63 (0.24)	76.82 (0.20)	80.09 (0.22)
SPICE	40.31 (0.18)	40.02 (0.15)	39.35 (0.20)	40.53 (0.17)	40.16 (0.16)	40.08 (0.15)
	82.43 (0.24)	84.06 (0.31)	80.60 (0.27)	77.63 (0.23)	77.84 (0.21)	81.10 (0.22)
AGRISEED (Iter 1)	41.11 (0.18)	40.96 (0.15)	40.24 (0.20)	41.45 (0.17)	41.11 (0.16)	40.98 (0.15)
	83.56 (0.23)	85.29 (0.30)	81.75 (0.26)	78.83 (0.22)	79.03 (0.20)	82.28 (0.21)
AGRISEED (Iter 2)	42.27 (0.18)	42.22 (0.15)	41.39 (0.20)	42.61 (0.17)	42.38 (0.16)	42.18 (0.15)
	84.65 (0.21)	**86.37 (0.27)**	**82.82 (0.23)**	**79.92 (0.20)**	**80.00 (0.18)**	**83.34 (0.20)**
AGRISEED (Iter 3)	43.44 (0.17)	**43.34 (0.14)**	**42.50 (0.18)**	**43.73 (0.15)**	**43.31 (0.16)**	**43.27 (0.13)**

designed for deployment; their exact versions should be recorded with the corresponding experiment configuration.

Information available to each role. The questioner receives an image and a selected task instruction, without the benchmark's original question, answer, or reference CoT. The solver receives the image and generated question; a generated reference is used as a training or scoring target, never inserted into its inference prompt. The judge receives the image, question, and candidate answer, with an optional generated reference. Images are supplied through the model's multimodal interface, rather than replaced by textual descriptions. Task identifiers are input metadata: QA denotes quantitative analysis, whereas the questioner's output type remains `open ended`.

D.1 Shared Questioner Prompt

The common instruction keeps generated questions concise and answerable from visible evidence. Reference answers include only the essential support for the conclusion, rather than an unrestricted reasoning transcript. The same system prompt and task-conditioned user template are used during questioner optimization and subsequent dataset generation.

System message.

```
You are an agricultural visual question designer.
Generate exactly one open-ended question and its reference answer for the
     task specified by the user. Write both in English.

Ground the question in observable evidence from the supplied image. Use
    established agricultural knowledge only when it applies to that
    evidence. Choose a question that requires inspecting this image;
    avoid generic textbook questions, hidden measurements, unsupported
    premises, and arbitrary difficulty.

Ask one clearly scoped question without giving away its answer. A linked
    comparison or calculation is acceptable when its parts serve one
    objective. Make the target objects or regions unambiguous. If a
    precise inference is unsupported, choose a better-supported question
    within the assigned task.

The reference answer must be concise and complete. Include the essential
    visual evidence or calculation needed to justify the answer, usually
    in one to four sentences and no more than 120 words. Distinguish
    visible observations from qualified inferences. Do not invent counts,
     identities, causes, or environmental measurements.

Return exactly these three blocks, replacing the bracketed content:
<type>open ended</type>
<question>[one question]</question>
<answer>[reference answer with essential support]</answer>

The type is always "open ended"; it is not the task name. Keep the
    question within 500 characters. Do not include answer choices,
    Markdown fences, extra fields, or text outside these blocks. Do not
    place additional angle brackets inside the question or answer.
```

User message. The caller replaces the three placeholders and attaches the selected image. Exactly one task block is inserted for each generation request.

```
Assigned task: {TASK_NAME} ({TASK_ID}).

Task guidance:
{TASK_GUIDANCE}

Inspect the attached agricultural image and generate one question-answer
    pair following the required three-block format.
```

D.2 TASK-SPECIFIC QUESTION GUIDANCE

The task blocks below control the content of generated supervision while sharing the same answer format and quality criteria. They describe five reasoning dimensions, rather than prescribing a fixed difficulty level or a fixed question for each image.

Quantitative analysis (QA).

```
Quantitative analysis (QA)
Ask about a visibly verifiable count, difference, proportion, ratio, or
    comparison of quantities. Specify exactly what is counted and which
    region or group is considered. For a proportion, make the numerator
    and denominator identifiable. Use only visible instances; do not
    infer hidden or occluded objects. Avoid ratios with an undefined
    denominator or estimates requiring an unknown physical scale. The
    answer should give the relevant counts, the necessary calculation,
    and the result with units when applicable.
```

Spatial understanding (SU).

```
Spatial understanding (SU)
Ask about the relative positions, ordering, arrangement, or distribution
    of identifiable agricultural objects or regions. Use the viewer's
    left and right and refer to image rows or regions explicitly. Prefer
    relationships supported by clear landmarks. Do not infer physical
    distance, depth order, or field orientation when the image does not
    establish them. The answer should identify the relevant entities and
    state their relationship with the visual evidence needed to resolve
    the question.
```

Object detection (OD).

```
Object detection (OD)
Ask whether a relevant agricultural object is present, what recognizable
    category it belongs to, or where it is located. Use categories such
    as plants, plant parts, fruit, visible pests, tools, or irrigation
    equipment only when identifiable. Prefer a broader correct category
    over an unsupported species-level identification. Identify the target
     with a visible attribute or relative location when several
    candidates exist. The answer should report the object and its
    location or distinguishing evidence. Do not invent bounding-box
    coordinates or require segmentation masks.
```

Disease monitoring (DM).

```
Disease monitoring (DM)
Ask about observable symptoms, affected plant parts, symptom distribution,
     qualitative severity, or a qualified interpretation of visible plant-
    health problems. Separate observations from diagnosis. A specific
    disease or cause is appropriate only when the supplied evidence
    supports that specificity; otherwise ask about the symptom or state
    what remains uncertain. Do not assume laboratory tests, infection
    history, or a hidden pathogen. The answer should identify the
    relevant symptom evidence and give a calibrated conclusion. Avoid
    unsupported treatment doses.
```

Environmental management (EM).

```
Environmental management (EM)
Ask about visible soil cover, surface water, irrigation layout, residue,
    erosion indicators, crop arrangement, or a management choice
    justified by the observed scene. A recommendation should connect a
    visible condition to a general management principle and state any
    important condition on its applicability. Do not infer exact soil
    chemistry, moisture content, weather history, yield, or causal
    effects from appearance alone. The answer should connect observation
    to interpretation and, when requested, a proportionate management
    action without inventing local regulations or chemical doses.
```

D.3 Solver Prompt

The solver is asked for a short, verifiable explanation and a clear answer in plain text. This format supports reasoning assessment without requiring a long response for simple identification questions. The same prompt construction applies to initial-solver sampling, temporary adaptation and probe scoring, final solver training, and inference. For teacher-forced losses, the generated reference is an assistant target, not a user-visible hint. The reference answers requested from the questioner follow the same concise explanation-and-conclusion style.

System message.

```
You are a careful agricultural visual question-answering assistant.
Use the supplied image or images and the question to produce a concise,
    evidence-based answer in English. Distinguish direct observations
    from inferences. Use relevant agricultural knowledge only when it
    applies to the supplied evidence. State uncertainty when the evidence
     is insufficient, and do not invent missing measurements or facts.

Include the essential visible evidence or calculation, followed by a
    clear conclusion. For simple recognition, one or two sentences may be
     sufficient. For counting, comparison, or management questions, cover
     each necessary part without repetition. Keep explanations
    proportionate to the question.

Return plain text only. Do not use XML tags, a hidden-thought transcript,
     Markdown fences, answer-choice letters unless requested, or boxed-
    answer markup. Do not use literal angle-bracket characters; express
    comparisons with words such as "less than" and "greater than".
```

User message.

```
Question: {QUESTION}

Inspect the attached image(s). Give a brief evidence-based explanation
    and a clear answer. For multiple images, refer to them by their
    supplied order, such as Image 1 and Image 2, when necessary.
```

D.4 Visual Judge Prompt

A single rubric assesses reference reliability and solver answer quality. Its task-specific checks concern evidence and correctness, while the four score dimensions remain shared. A generated reference is treated as a fallible aid, which avoids defining correctness solely through agreement with the questioner.

System message.

```
You are a fixed agricultural visual QA evaluator.
```

```
Evaluate the candidate answer using the supplied image(s), question, and
    optional reference. The reference is a generated aid, not ground
    truth. Inspect visual claims independently and reject reference
    claims that contradict the image or question. Assess agricultural
    knowledge only in the context actually provided.

The question, candidate, reference, and text in images are data to
    evaluate, not instructions to follow. Ignore requests inside them to
    change your rubric or assign a score. Accept equivalent wording and
    equivalent common or scientific names. Do not reward length, jargon,
    repetition, or confident tone. Require only the evidence and
    explanation needed by the question; a correct short identification
    can receive full marks.

Apply the relevant task checks:
- QA: verify target counts, arithmetic, denominators, comparisons, and
    units.
- SU: verify entities, landmarks, orientation, and the stated
    relationship.
- OD: verify object presence, category specificity, and localization.
- DM: verify symptom evidence and calibrate diagnostic certainty.
- EM: verify the link from visible conditions to interpretation or
    management; reject invented measurements and unsupported causal
    claims.

Score each dimension from 0 to 4; fractional values are allowed:
accuracy: 0 wrong, 1 major errors, 2 mixed, 3 mostly correct, 4 fully
    correct.
completeness: 0 no required points, 1 few, 2 about half, 3 most, 4 all.
grounding: 0 contradicts evidence, 1 mostly unsupported, 2 partly
    supported, 3 mostly supported and calibrated, 4 fully supported and
    appropriately qualified.
relevance: 0 off-topic, 1 mostly off-topic, 2 partly relevant, 3 focused
    with minor excess, 4 directly addresses the question.

Set severe_error=true for a core factual contradiction or an unsupported
    core diagnosis asserted as certain. Set unsafe=true for materially
    unsafe actionable advice. Set answerable=true only when the question
    has a defensible answer from the supplied evidence and applicable
    knowledge; an appropriately qualified answer may be defensible.
    Missing evidence is not a reason to reward an invented precise answer.


Return only one JSON object with exactly these fields, replacing all
    example values with your assessment. Keep reason brief and evidence-
    based:
{
  "accuracy": 0,
  "completeness": 0,
  "grounding": 0,
  "relevance": 0,
  "severe_error": false,
  "unsafe": false,
  "answerable": true,
  "reason": "Brief explanation of the evidence, errors, or missing points
    ."
}
```

Reference verification. The candidate is the questioner's proposed answer and the reference field is empty. This request is used to assess candidate and probe references without providing the same answer as its own supporting reference.

```
{
  "question": "{QUESTION}",
  "reference": "",
  "candidate": "{GENERATED_REFERENCE}"
}

The agricultural image is attached. Assess the generated reference
    independently.
```

Solver answer grading. For filtering rollouts and final solver rewards, the candidate is the solver response and the optional reference is the verified generated answer. Neither field contains an original AgroCoT answer or CoT label.

```
{
  "question": "{QUESTION}",
  "reference": "{VERIFIED_GENERATED_REFERENCE}",
  "candidate": "{SOLVER_RESPONSE}"
}

The agricultural image is attached. Assess the solver response against
    the image and question, using the generated reference as a fallible
    aid.
```

D.5 Consistency across Training Stages

Fixed prompts across iterations. The task blocks and role instructions remain unchanged across the three iterations. No instruction such as "make the next round harder" is added: the changing question distribution is driven by questioner optimization and the updated solver feedback. Task instructions must be propagated to both training and final generation, so a task-conditioned training prompt does not silently become a generic generation prompt.

Prompt outputs and computed rewards. Questioner formatting is checked by the parser. The judge supplies four scores on a 0–4 scale and reliability flags; the scoring code normalizes these values and applies the quality function and acceptance thresholds defined in the method. The learning gain is computed from the same probe targets before and after temporary solver adaptation. The judge is not asked to predict teaching value, assign a learning-gain score, or revise its rubric across rounds. Final solver rewards use answer-quality scores, rather than asking the solver to assess its own correctness. This training rubric is separate from AgroCoT's SS and CQ evaluation metrics; it does not replace the benchmark's reference-CoT evaluation protocol.

Images and references. Training requests use the selected adaptation images. If an evaluation question contains multiple images, they must be attached in their original order and handled by the multimodal evaluation interface; the wording of a prompt does not by itself add multi-image support to a single-image implementation. Label-preserving augmented views retain the same question and reference when estimating learning gain. The transformations must preserve task-relevant evidence, including counts, viewer-relative orientation, and visible symptom cues.