AGROGROUND: MULTI-GRANULARITY GROUNDED RECOGNITION IN AGRICULTURE

Anonymous authors
Paper under double-blind review

ABSTRACT

Agricultural visual models are generally evaluated for either recognition or localization, but reliable diagnosis requires identifying what is present and localizing the visual evidence that supports the prediction. Agricultural visual question answering (VQA) datasets contain rich semantic labels but often lack annotations linking them to image regions. Adding these annotations by hand is costly at scale. We introduce AgroGround, a large-scale dataset for grounded agricultural recognition, where models identify plant diseases and other agricultural targets and localize their corresponding image regions. We develop an automated annotation pipeline that uses semantic labels from eight existing agricultural VQA datasets to annotate both fine-grained regions such as disease lesions and whole objects, producing 794,850 instruction examples. Healthy images provide negative supervision for disease-related queries, teaching the model to return empty predictions. We train a shared vision-language model through supervised fine-tuning, combining known-target grounding instructions with instructions that require both target recognition and localization. We evaluate predicted identities, regions, their joint correctness, and healthy-image abstention on 1,480 human-verified images disjoint from all training data by file path and perceptual hash. Grounding-only fine-tuning reduces recognition accuracy from 51.8% to 29.1%, and adding recognition-and-localization instructions raises it to 72.6%. With positive images and spatial annotations held fixed, combining the two instruction formats raises joint accuracy from 19.2% to 43.3% with comparable grounding performance. Healthy negatives raise abstention on healthy images to 95.0%, and reinforcement learning improves lesion-level grounding. The resulting 2B model achieves higher grounding F1 than its annotation teacher on both our benchmark and the external PlantSeg test set. AgroGround establishes a benchmark for grounded agricultural recognition, measuring joint correctness of target identity and localization, as well as abstention on healthy images.

1 INTRODUCTION

Automated plant disease assessment is only actionable when a model does two things: identify the disease and localize the supporting image regions, so growers can verify predictions and plan treatment. Current systems typically do one or the other: disease classifiers and agricultural visual question answering (VQA) systems name the disease without localizing the evidence (Mohanty et al., 2016; Liu et al., 2024b), while detection, segmentation, and grounding models localize symptoms but either do not name the disease (Dubois et al., 2026) or require the query to specify it (Li et al., 2026a; Haghighat et al., 2026). Performing both jointly has received little attention.

We address this gap with **AgroGround**, which couples large-scale spatial supervision with a human-verified benchmark and a shared model for recognition, localization, and abstention. First, we construct a dataset and benchmark for *grounded agricultural recognition*: given an image and a query, a model must state the target and localize the corresponding regions, or return an empty output when the queried target is absent (Figure 2). Diseases are the primary focus, but pests and other agricultural targets are also included. Second, we train a shared vision-language model on grounding and recognition jointly.

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.

Grounded recognition needs semantic labels paired with spatial annotations, which few agricultural resources offer at scale: VQA datasets carry rich labels but rarely image regions (Liu et al., 2024b; Shinoda et al., 2025; Nguyen Quoc et al., 2026), while human-annotated box or mask sets are small and fixed-class (e.g., PlantSeg contains 11,458 images of 115 diseases; Wei et al., 2026; Singh et al., 2020; Wu et al., 2019).

We therefore build AgroGround with an automated pipeline that converts the labels of eight agricultural VQA datasets into spatial supervision, following open-vocabulary auto-labeling practice (Peng et al., 2024; Xiao et al., 2024). Of 1,274,601 samples, 615,485 name a visually localizable target. The crop and target in each answer prompt GroundingDINO (Liu et al., 2024a), SAM2 (Ravi et al., 2025) converts its boxes to masks, and geometry checks remove unreliable predictions and assign the supervision level. The pipeline yields 794,850 instruction examples across the two granularities, one per supported level.

Models trained only on positive examples rarely abstain, returning regions even when the queried disease is absent (abstention at most 0.2% without negatives, Table 3), so we retain healthy images as negatives for disease-related queries, with the empty output as the answer. The automatic boxes are treated as noisy and used only for training. Evaluation uses a human-verified benchmark of 1,480 images (508 lesion-level, 467 object-level, 505 healthy), with no overlap detected against training data under file-path and perceptual-hash audits, scoring the predicted target and regions, their joint correctness, and abstention on healthy images, with recognition evaluated under a closed-set candidate protocol.

Spatial annotations alone do not teach the joint task: known-target instructions contain the target name, so a model can localize without recognizing. We therefore fine-tune one model on a mixture of known-target and *recognition-and-localization* instructions, the latter listing four candidates to select among and localize, both formats sharing images and boxes. Reinforcement learning (RL) with grounded rewards on the automatic boxes then targets the main remaining failure, missed lesions on many-instance images.

Instruction format determines which capabilities fine-tuning preserves: grounding-only training reduces zero-shot recognition from 51.8% to 29.1%, whereas adding recognition-and-localization instructions raises it to 72.6% and joint accuracy from 19.2% to 43.3%. Healthy negatives raise abstention on healthy images to 95.0% at 4.6% false abstention, and RL raises lesion-level F1 from 0.392 to 0.439. After RL, the 2B model exceeds its annotation teacher in grounding F1 on our benchmark and on PlantSeg. The three capabilities are coupled: improving one can degrade another, so they should be trained and evaluated jointly.

Our contributions are:

- An automated annotation pipeline and the large-scale dataset built with it, converting the labels of eight agricultural VQA datasets into 794,850 lesion- and object-level instruction examples, with healthy images as negatives for disease-related queries.
- A single vision-language model trained on both instruction formats, improving joint accuracy at comparable grounding. Minimal healthy negatives enable precise healthy-image abstention without compromising recognition, and grounded RL lifts lesion-level localization beyond teacher annotations.
- A 1,480-image human-verified benchmark, audited for training overlap, measuring identity, localization, joint correctness, and abstention, with controlled comparisons and external evaluation on PlantSeg.

2 Related Work

Agricultural vision–language data. Large agricultural multimodal corpora (Liu et al., 2024b; Shinoda et al., 2025; Gauba et al., 2025; Dongre et al., 2025; Wen et al., 2025; Li et al., 2025; Nguyen Quoc et al., 2026; Boudiaf et al., 2026; Sakib et al., 2026; Mahmood et al., 2026) carry text supervision but almost no box- or mask-level ground truth. AgroGround converts this text supervision into spatial supervision. Human-annotated localization sets are fixed-class and smaller (PlantSeg, PlantDoc, IP102), with PlantSeg as our expert-label comparison (Appendix N). AgriDoctor (Zhang et al., 2026) and AgriMM (Wang & Liu, 2026) carry manual boxes, but in neither does language select the

target, and language-conditioned agricultural grounding data exists for pests (Yang et al., 2026b), not disease. Closest to us, AgroVG (Li et al., 2026a) provides 10,071 evaluation-only image–query pairs across six target families, and gRef-CW (Haghighat et al., 2026) adds existence-aware grounding for crops and weeds (single-source crop and weed targets). AgroGround complements both with a training corpus, human-verified evaluation, disease-level targets, and grounded recognition.

Automatic grounding supervision. Auto-labeling text with open-vocabulary detectors underlies GLIP (Li et al., 2022), Kosmos-2 (Peng et al., 2024), and Florence-2 (Xiao et al., 2024), and has been applied to industrial detection (Xin et al., 2024), tomato leaf disease (Li et al., 2024), and medicine (Xie et al., 2025). Our pipeline differs in labeling a curated 6,038-label taxonomy with crop-aware prompts, converting VQA answers rather than captions, and validating on a human-verified, duplicate-audited evaluation. Confident GroundingDINO+SAM false positives typically occupy large areas and fall to relative-size filtering, motivating our geometry filters (Mumuni & Mumuni, 2024). Text-prompted leaf-symptom masks range from IoU 0.19 to 0.46 across prompts on their primary crop (Dubois et al., 2026).

Open-vocabulary detection and VLM grounding. GroundingDINO (Liu et al., 2024a), OWLv2 (Minderer et al., 2023), and SAM/SAM2/SAM3 (Kirillov et al., 2023; Ravi et al., 2025; Carion et al., 2026) are trained on general imagery, far from agricultural targets. Modern VLMs emit box coordinates directly (Qwen Team, 2025), RL fine-tuning sharpens them (Zhan et al., 2026; Shao et al., 2024), reason-then-localize training extends reasoning segmentation (Lai et al., 2024; Liu et al., 2025a; 2026; Hegde et al., 2026), and IoU-style verifiable rewards are now standard (Liu et al., 2025b; Shen et al., 2025; Yu et al., 2025; Peh et al., 2026). Our reward is set-level (matched instance-F1, region IoU, recall, recognition, and two-sided abstention). We adopt this recipe and study how rewards computed from automatically generated boxes affect localization, recognition, and abstention. Concurrent WeedExpert-R1 (Yang et al., 2026c) pairs SFT with GRPO on Qwen3-VL but grounds named weed species only and reports neither a target-absent condition nor a recognition or abstention reward.

Grounded diagnosis and existence awareness. UniBiomed (Wu et al., 2026) and 3DReasonKnee (Sambara et al., 2026) require diagnosis and localization in one pass but score them separately, and XBench (Luo et al., 2026) finds recognition-strong, grounding-weak VLMs. We introduce the agricultural counterpart and observe a within-model dissociation in the opposite direction. Grounding fine-tuning is known to cost other abilities (Wu et al., 2025; 2024), and such loss is repairable from the data side (Chen et al., 2025; Li et al., 2026b). In our setting, rewriting the instructions alone, with answers unchanged, restores recognition. On existence awareness, GroundingME (Li et al., 2026c) evaluates 25 state-of-the-art MLLMs and finds most at 0% on rejection. At 1:8, nearest our share, rejection rises from 30.5 to 83.5 with positives unharmed, falling (88.2 to 83.1) at 2:1, though their out-of-domain grounding degrades at every ratio (38.8% to at most 33.0%). Our own out-of-domain transfer shows no such cost (PlantSeg F1 0.231 with and without negatives, Appendix O). Our regime instantiates generalized referring (Liu et al., 2023; He et al., 2023): gRefCOCO's N-acc. and T-acc. are our abstention and one minus our false abstention. Concurrent RC-GRPO (Yang et al., 2026a) rewards two-sided refusal but on general-domain data with no recognition term.

3 The AgroGround Dataset and Benchmark

3.1 Task Definition

Given an image and a text instruction, a model outputs a list of bounding boxes, each labeled with a target name, or an empty list when the queried target is absent. AgroGround poses the task in two instruction formats that share this output. A *known-target grounding* instruction names the target and asks where it occurs. A *recognition-and-localization* instruction instead lists four candidate targets and asks the model to identify the one that is present and localize its evidence, following closed-set formulations of grounded diagnosis (Wu et al., 2026; Sambara et al., 2026). This format requires the model to determine the target itself. The candidate list contains the true label and three distractors, drawn preferentially from diseases of the same crop, in random order. Healthy images are queried in either format with plausible diseases of the same crop, and the correct output is an empty list. We refer to evaluation with the two formats as the *disease-given* and *candidate* protocols (Figure 6). Each instruction also specifies a spatial granularity: individual lesions (*lesion level*) or affected leaves, fruits, and plants (*object level*). The distinction is needed because a whole-leaf box

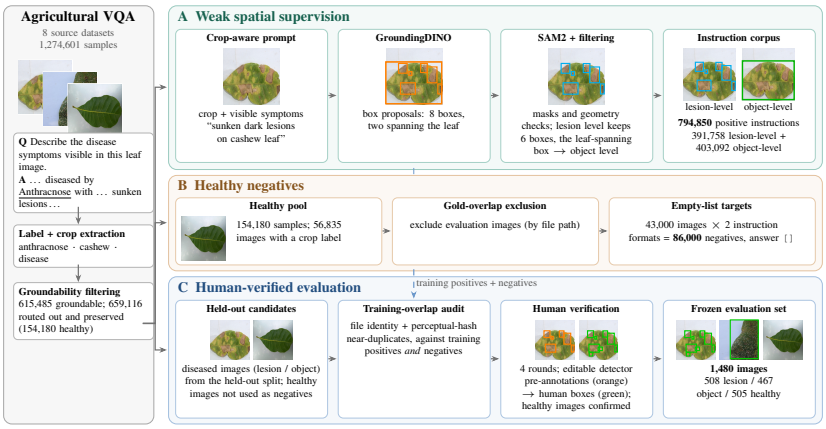

Figure 1: **Data construction in AgroGround.** Left: target labels and host crops are extracted from the answers of eight agricultural VQA datasets, and samples that name a visually localizable target are retained as groundable (§3.2). (A) For each groundable sample, a crop-aware prompt drives GroundingDINO (Liu et al., 2024a). SAM2 (Ravi et al., 2025) masks and geometry checks filter the proposals, and the retained boxes are assigned to lesion-level or object-level instructions. (B) Healthy images, excluding benchmark images, provide negatives in both instruction formats with an empty target (§4.3). (C) Held-out candidates are audited against training positives and negatives by file identity and perceptual hash, and detector pre-annotations (orange) are corrected into human-verified boxes (green) (§3.3). All images, prompts, and boxes shown are real. The running example is a benchmark image of cashew anthracnose, and its automatic boxes are shown for illustration only and were not used for training.

can describe a diffuse symptom but does not localize individual lesions, and the annotation level measurably affects plant disease detection (Dong et al., 2022). Targets are mainly diseases but also include pests, weeds, and other visible agricultural targets. For brevity, we refer to them as diseases.

3.2 Automatic Annotation of Training Data

AgroGround derives its training supervision from eight agricultural VQA datasets comprising 1,274,601 question–answer samples (Table 4). The pipeline extracts a target label from each answer, generates boxes for this target with an open-vocabulary detector, and assigns each annotated sample to one or both granularities (Figure 1).

Label extraction. Because not every VQA answer names something that can be localized, source-specific handlers classify each answer by the type of target it names, such as a disease, pest, management action, or healthy plant, and extract the target label and host crop. Answers that name a visually localizable target are retained as *groundable*. Management actions, causes, aerial field-level labels, plant-identification questions in CDDM, and non-agricultural or overly generic labels are excluded, and the 154,180 healthy samples are kept in a separate pool for negative supervision (§4.3). This yields 615,485 groundable samples, 97.6% of which name a disease. Labels are normalized to a 6,038-entry taxonomy of diseases, pests, and weeds, each with a visual-symptom description. Normalization merges synonyms and replaces a label's default host with the crop shown in the image (Appendix A).

Box generation. A disease name specifies neither host crop nor symptom, so each sample's crop, normalized label, and symptom description form a prompt for GroundingDINO (Liu et al., 2024a), and SAM2 (Ravi et al., 2025) converts the proposed boxes into masks. A box-quality audit found detector confidence only weakly related to box quality, so boxes are filtered by geometry rather than by score: degenerate boxes, boxes covering nearly the full image, and boxes with empty masks are removed (Mumuni & Mumuni, 2024). The detection thresholds were fixed before any evaluation data were collected (Appendix A).

The retained boxes also determine granularity. A lesion-level example keeps small, localized boxes, after boxes covering more than half of the image are removed and non-maximum suppression is

applied, while an object-level example keeps a single box spanning the affected leaf, fruit, or plant. An image that supports both granularities contributes one example at each level, so diffuse diseases, which typically yield whole-leaf boxes, contribute object-level rather than lesion-level supervision.

Resulting corpus. Of the 614,765 readable groundable samples, 579,260 (94.2%) receive at least one valid box, yielding 794,850 instruction examples: 391,758 at the lesion level and 403,092 at the object level. The 171,580 training image files contain 104,150 perceptually unique images, so example counts measure supervision volume rather than visual diversity. In a stratified 100-image audit, all boxes were correct in 62, and most remaining errors were granularity or instance coverage (Appendix M). These boxes serve only for training.

3.3 Human-Verified Benchmark

Construction and disjointness. Headline claims require ground truth independent of the automatic annotator and verifiably disjoint from training. Candidates are drawn from the held-out test split and pass two rules: no shared file with the training split, and no perceptual near-duplicate of any training image, positive or negative, at Hamming distance ≤ 6 on a 64-bit perceptual hash with pixel-level verification of flagged pairs. The second rule matters: an earlier 1,530-image version built under the file-path rule alone failed it, and 564 pixel-verified training copies, 8 ambiguous pairs, 34 within-set repeats, and 202 healthy near-copies were removed (Appendix B). The 722 surviving images (537 positives, 185 healthy) were joined by a fourth round applying both rules (438 of 452 new positives kept, 320 confirmed healthy). All four rounds followed one protocol (Appendix C): editable detector pre-annotations, informed labeling, tight boxes, exhaustive instance coverage, granularity classification, and removal or flagging of unverifiable images, and healthy images were visually confirmed. The final benchmark contains **1,480 images: 508 lesion-level, 467 object-level, 505 healthy**, with 3,463 human boxes (5.49 per lesion-level image and 180 dense images, mean 12.2). Composition reflects deduplication survival (Appendix B): 463 of 975 positives are in-the-wild photographs from MIRAGE and AgroBench, and 492 of 505 healthy images come from the three leaf-centric sources. Positives are 62% diseases (607). The rest name pests, weeds, and other visible agricultural targets (Appendix B). Ground truth was frozen before model comparison and never revised toward model outputs. The corpus's sample-level validation split was not used for model selection and shares photographs with the training partition, so it measures fit rather than generalization (Appendix D). No benchmark image was used in training or in computing any validation signal. Two reporting choices were made after seeing benchmark results, the 12% negative share and the step-170 RL checkpoint, and both alternatives are also reported (Table 3; Appendix K).

3.4 Evaluation Protocols and Metrics

Metrics. Per-instance: greedy one-to-one matching at IoU 0.5, yielding set precision, recall, and F1 (AgroVG-style set matching, Li et al., 2026a, rules in Appendix D). Region-level: IoU between the unions of predicted and ground-truth boxes. Abstention: correctly empty output on the 505 healthy images. False abstention: empty output on diseased images. Recognition is scored on the 975 positives (chance 25%), with the position distribution of the model's picks as a probe for prompt-order bias, and joint accuracy is correct disease *and* region IoU ≥ 0.5. Intervals are 95% percentile bootstraps over images (1,000 resamples, fixed seed per statistic). Paired differences report the observed difference between models on the same images with a seeded bootstrap interval.

4 Learning Grounded Recognition

Figure 2 summarizes our approach. A single model with one output format (§4.1) is trained with supervision in which each component targets one capability: mixed instruction formats for recognition (§4.2), healthy negatives for abstention (§4.3), and reinforcement learning with grounded rewards for instance coverage (§4.4).

4.1 Model and Output Format

We fine-tune Qwen3-VL-Instruct (Qwen Team, 2025) with the next-token prediction loss. Every answer is text: a JSON list of boxes labeled with the target name, and an empty list denotes abstention.

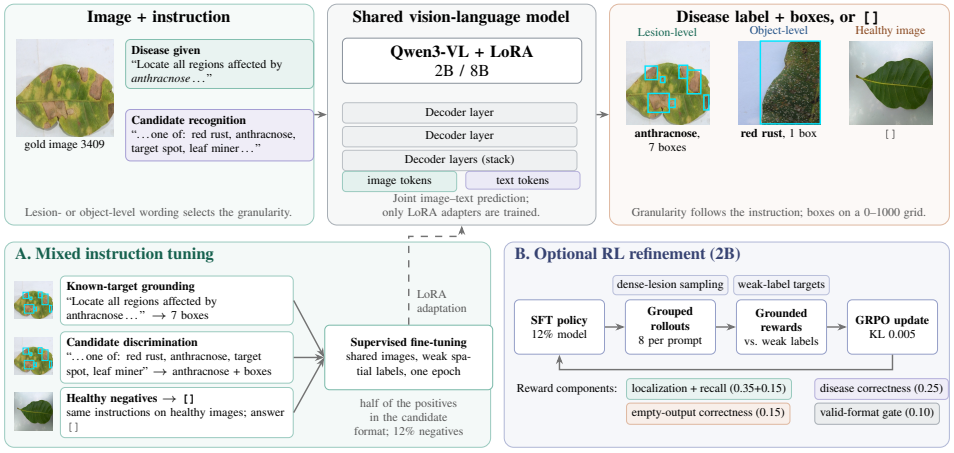

Figure 2: **Overview of the AgroGround model and training.** Top: a shared vision-language model takes an image and an instruction, either known-target grounding or recognition and localization, and outputs the target name with boxes at the requested granularity, or an empty list when the queried target is absent. (A) Supervised fine-tuning mixes the two instruction formats with healthy negatives whose target is an empty list (§4.2–4.3). (B) Reinforcement learning starts from the supervised model and optimizes a grounded reward for localization, recall, recognition, abstention, and valid format (§4.4). Example outputs are predictions of the supervised 2B model on benchmark images.

Because recognition, localization, and abstention share this output, one model serves both formats and granularities with no task-specific heads.

4.2 Mixing Grounding and Recognition Instructions

Known-target grounding does not require recognition: the instruction already names the target and the answer label repeats it, so a model can learn to localize without determining what the image shows. The *mixed* configuration therefore renders a random half of the positive examples in the recognition-and-localization format and the rest in the known-target format. Because both formats share images and boxes, the added format supervises only the choice of target. Candidate lists follow §3.1: same-crop distractors curb host-crop shortcuts, and randomized answer positions counter candidate-order sensitivity (Zheng et al., 2024; Zong et al., 2024; Xue et al., 2024).

4.3 Healthy-Negative Supervision

Positive examples never demonstrate an empty answer, so a model trained only on them has no signal for abstaining. The *mixed* + *healthy* configuration therefore adds negatives from the healthy pool: each selected healthy image is queried in both formats with plausible diseases of the same crop, and the target is an empty list.

4.4 Reinforcement Learning with Grounded Rewards

Supervised fine-tuning maximizes the likelihood of one reference sequence and gives no explicit credit for recovering every instance, which limits recall on images with many lesions. From the mixed + healthy model, we therefore apply GRPO (Shao et al., 2024) with the GSPO loss variant (Zheng et al., 2025), following the grounding recipe of Zhan et al. (2026). Sampled responses are scored against the automatically generated boxes with a five-term reward $R(y) = \mathbf{1}[\text{valid format}] \cdot (0.10 + 0.35\,\ell + 0.15\,m + 0.25\,r + 0.15\,a)$, where ℓ is the mean of instance F1 and region IoU, m is instance recall, r credits the correct candidate (0.25 if listed but incorrect), and a rewards an empty output on healthy and a non-empty output on diseased images. The format gate assigns zero reward to unparseable outputs, the recall term m adds credit for instance coverage, and r and a keep recognition and abstention in the objective. Terms that do not apply to a prompt type are dropped

or redistributed (Appendix J). Rollout prompts are drawn from the training split and oversample images with many automatically generated boxes, concentrating learning on instance coverage.

5 EXPERIMENTS

5.1 EXPERIMENTAL SETUP AND DATASET ANALYSIS

Training setup. Table 1 contrasts the training corpus with the human-verified benchmark. Algorithm 1 summarizes both training stages. We fine-tune Qwen3-VL-2B- and 8B-Instruct (Qwen Team, 2025) with LoRA (Hu et al., 2022) and the official toolchain (Appendix J), training images capped at 50,176 pixels. No hyperparameter was searched on our data, and evaluation applies no pixel cap, so every fine-tuned model is evaluated above its training resolution (Appendix D). Answers are JSON boxes in the model's native 0–1000 space with the lowercased disease name as label. Examples are split 80/10/10 by source sample (635,977 / 79,464 / 79,409, Section 3.3).

	Training	Benchmark
Annotation	automatic	human-verified
Image files	171,580	1,480
Lesion level	391,758 ex.	508 img.
Object level	403,092 ex.	467 img.
Healthy images	43,000	505
Disease targets	97.6%	62.3%
MIRAGE + AgroBench	2.2%	47.5%

Table 1: Training corpus and benchmark. Percentages are over groundable samples (training) and positive images (benchmark).

Healthy negatives. We add negatives from the preserved healthy stream. The pool excluded benchmark images by file path only, and the benchmark was perceptual-hash filtered against it, so no healthy benchmark image has a near-duplicate among the negatives (Appendix B). Each selected image contributes two negatives: candidate-format (four plausible same-crop candidates, answer an empty list) and disease-given (a plausible same-crop disease query, empty list). Two rules prevent shortcuts (Geirhos et al., 2020): identical label casing everywhere, and every phrasing carries both answer types, so no surface feature predicts the empty answer. Both shares were specified in advance and both are reported (21,500 and 43,000 healthy images, two negatives each, Appendix H). All else matches the mixed model.

Baselines. GroundingDINO under three prompt configurations, two zero-shot 7B reasoning-grounding systems, and the zero-shot 2B and 8B bases (adaptations in Appendix G).

5.2 MAIN RESULTS AND COMPARATIVE ANALYSIS

Student versus teacher. We score the teacher three ways (Table 2): raw names, the annotation-time taxonomy prompts, and those prompts plus the pipeline's geometry filters, the configuration that labeled the corpus. All three land within 0.009 (Appendix I), so neither better prompts nor hand-built filters rescue the detector. Supervised fine-tuning alone already exceeds the annotator configuration, a paired +0.048 [+0.021, +0.077], while against raw prompts the difference is +0.039 [−0.001, +0.079] and is not resolved. The full pipeline is decisive: after RL the student reaches 0.477 [0.448, 0.509], a paired +0.085 [+0.057, +0.113] over the annotator configuration, +0.076 [+0.038, +0.116] over raw prompts and +0.082 [+0.049, +0.115] over taxonomy prompts (Table 2). Every AgroGround model also exceeds VisionReasoner-7B (+0.057 to +0.108) and Seg-Zero-7B (+0.179 to +0.230), all paired intervals excluding zero. Failure profiles differ (Figure 4, Appendix I): GDINO *over-covers* while the fine-tuned model *under-enumerates* (1.7–2.2 boxes against 5.5 ground-truth instances per lesion-level image). Instance recovery, not box tightness, drives the margin, and the gains concentrate on dense images while the zero-shot 8B leads on non-dense ones (Appendix E). The 7B systems localize regions but not instances. The zero-shot 8B base has the best region IoU (0.644), abstaining unprompted on 31% of healthy images. Apart from that and a filter artifact (the annotator configuration's 20.2%), abstention is absent from every model trained or thresholded on positives, addressed in §4.3.

Grounding-only training erases zero-shot recognition. The zero-shot 2B model recognizes the disease in 51.8% of cases (8B: 64.4%). The grounding-only model (Table 3), fine-tuned on the same base, scores **29.1%**, barely above the 25% chance level. The position probe exposes the mechanism: 88.8% of its in-list picks are the first-listed candidate (Figure 3), and 77 answers name a disease not in the list. Grounding-only supervision, whose answer label always restates the disease in the instruction, teaches the model to echo the prompt rather than read the image. The bias is *induced* by

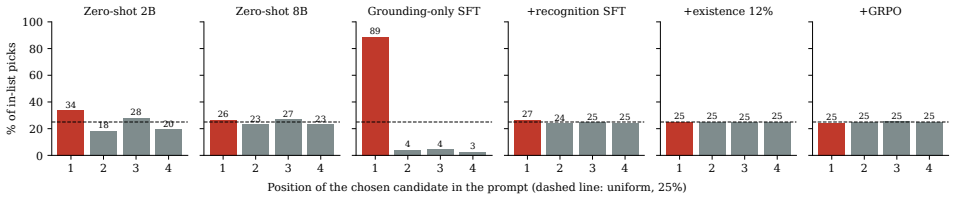

Figure 3: Grounding-only training echoes the first-listed candidate, while recognition training with randomized answer positions yields a uniform distribution. The model decides from image content, not prompt order.

	Localization		Recognition		Abstention		External
Model	F1 [95% CI] ↑	Lesion F1 ↑	Acc. (%) ↑	Joint (%) ↑	Healthy (%) ↑	False (%) ↓	PlantSeg F1 ↑
GroundingDINO (Liu et al., 2024a)							
W. raw prompts	0.401 [0.368, 0.434]	0.323	27.6†	17.5†	0.0	0.1	0.182
W. taxonomy prompts	0.395 [0.374, 0.416]	0.358	–	–	0.0	0.0	0.183
W. annotator configuration	0.392 [0.363, 0.425]	0.282	–	–	20.2	0.5	0.161
VisionReasoner-7B (Liu et al., 2026)	0.369 [0.340, 0.401]	0.302	40.6†	26.8†	0.2	0.5	–
Seg-Zero-7B (Liu et al., 2025a)	0.247 [0.223, 0.276]	0.124	–	–	0.0	0.0	–
Qwen3-VL-2B, zero-shot	0.320 [0.288, 0.357]	0.205	51.8	34.5	1.2	0.0	0.164
Qwen3-VL-8B, zero-shot	0.380 [0.347, 0.418]	0.294	64.4	44.6	30.9	2.5	0.215
Ours 2B, mixed + healthy	0.440 [0.409, 0.474]	0.392	74.7	44.8	95.0	4.6	0.231
Ours 2B, mixed + healthy + RL	**0.477** [0.448, 0.509]	**0.439**	69.2	43.1	87.7	1.0	**0.282**
Ours 8B, mixed + healthy	0.429 [0.401, 0.461]	0.369	**78.2**	**48.6**	**96.0**	2.9	0.237

Table 2: **Comparison with existing methods.** Localization and abstention use the disease-given protocol, and recognition uses the candidate protocol. †Protocol adaptations in Appendix G. –: not supported or not evaluated. Complete results are in Appendix H.

supervision whose own task lists no candidates (§4.2). Its localization is intact: with predicted labels ignored it reaches F1 0.441 against the mixed model's 0.432 (Appendix E). It boxes the diseased region while naming it wrong, a localization-without-recognition dissociation within a single model.

Recognition supervision via instruction format alone. Keeping the training recipe fixed and changing only the instruction distribution, we convert 50% of training examples to the candidate format (four crop-matched candidates, disease not given, true answer at a random position, the direct countermeasure to echo bias) while the rest keep the disease-given format. Answers, split, and hyperparameters are unchanged. Recognition rises to **72.6%** [70.0, 75.4] from 29.1% (Table 3). The mechanism changes, not just the score: the position distribution becomes uniform (26.7% first-pick) and off-list answers drop from 77 to 7. Joint accuracy more than doubles (19.2% → 43.3%). Blanked-image and tile-shuffling controls show the signal is local lesion texture (Appendix F). Recognition supervision preserves grounding on the disease-given task (F1 0.435 vs. 0.426, paired difference +0.009, 95% CI [−0.017, +0.035]), while increasing lesion enumeration (2.00 vs. 1.73 boxes per lesion-level image), an increase whose effect on F1 is not statistically resolved. Per source, base-model knowledge and in-domain supervision are complementary (Appendix I). Fine-tuning the 8B with the same recipe combines the two: recognition 78.2% [75.4, 80.7] (paired +3.5 points over the 2B, CI [+0.7, +6.1]), joint accuracy 48.6% (paired +3.8 points over the 2B, CI [+1.0, +6.5]), and candidate-protocol abstention 94.7%, the highest in Table 9. Per source, it keeps the base's MIRAGE strength while gaining the in-domain expertise, though not the base's AgroBench edge (Appendix I). Recognition and joint accuracy are higher for the 8B, while lesion enumeration is not, and two model sizes do not establish a trend (Table 2). What remains at 0% for models trained without negatives is abstention: no training example shows a "none" answer. Open-set naming with no candidate list remains near zero for every model (Appendix L), hence the closed-set formulation.

Negative supervision is cheap: at 12% negatives the model rejects 95.0% of healthy images under a crop-matched disease query at 4.6% false abstention, while recognition, grounding F1, and joint accuracy are unchanged or slightly improved and box emission rises (Tables 2 and 12). Doubling the share from 6% changes no metric beyond seed variance (Appendix N). Region IoU appears to fall, but most of the drop is mechanical, from empty predictions scoring zero (Appendix D). On shared answered images the cost attributable to the boxes is 0.014. Our 1:7.4 point sits in the low-ratio

Configuration	F1 ↑	Lesion F1 ↑	Recog. (%) ↑	Pos.-1 (%)	Joint (%) ↑	Healthy abst. (%) ↑	False abst. (%) ↓
(a) Training components							
Zero-shot Qwen3-VL-2B	0.320	0.205	51.8	33.7	34.5	1.2	0.0
Grounding only	0.426	0.366	29.1	88.8	19.2	0.0	0.0
Mixed	0.435	0.377	72.6	26.7	43.3	0.2	0.0
Mixed + healthy (6%)	0.433	–	73.0	–	44.3	93.3	5.2
Mixed + healthy (12%)	0.440	0.392	74.7	25.1	44.8	95.0	4.6
Mixed + healthy + RL	0.477	0.439	69.2	24.5	43.1	87.7	1.0
(b) RL design, one epoch							
Uniform rollout sampling	0.440	0.409	70.2	–	40.9	88.1	4.9
Recognition-weighted reward	0.480	0.436	70.2	–	44.1	84.2	3.0

Table 3: **Ablation on the 2B model.** (a) Supervised configurations are trained from the base model. RL starts from mixed + healthy (12%). (b) One-epoch RL variants. The main RL model uses step 170 (about 1.2 epochs). Pos.-1: share of in-list answers that select the first candidate.

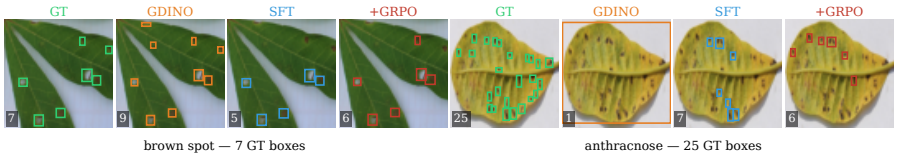

Figure 4: Disease-given grounding on two dense benchmark images (rows 1 and 6 of Figure 7, seed 3). Numbers are box counts.

regime of Li et al. (2026c) (§2). The 12% model, chosen after evaluating both shares, is the main model and the RL initialization.

RL delivers its target. Relative to its initialization, grounding F1 rises from 0.440 to **0.477** (paired difference +0.037, 95% CI [+0.020, +0.052]), recall from 0.320 to 0.362 (paired +0.042, 95% CI [+0.029, +0.057]) at essentially unchanged precision, lesion-level F1 from 0.392 to 0.439, and false abstention falls from 4.6% to 1.0% (Tables 2 and 8). The cost is recognition: 74.7% → 69.2% (paired −5.4 points, 95% CI [−7.5, −3.2]), with joint accuracy unchanged within noise (44.8% → 43.1%, paired −1.7, CI [−4.0, +0.6]) and abstention on healthy images lower under the disease-given query (paired −7.3 [−10.1, −4.8]).

On PlantSeg the RL model improves on its initialization by +0.051 F1 (CI [+0.042, +0.059]) and is the best model there. Under pseudo-label rewards, RL buys better lesion enumeration and far fewer false abstentions at a 5.4-point recognition cost. On wrong-target queries abstention is health detection rather than named-target verification (Appendix F).

5.3 Ablations and Qualitative Results

Table 3 isolates each training component (a) and the RL design choices (b): reward re-weighting does not remove the recognition cost, a uniform rollout mix forfeits the grounding gain, and the final checkpoint shows the over-optimization signature (Appendix K). Figure 4 compares predictions with the human-verified reference (full grid in Appendix P).

6 Discussion and Conclusion

Limitations. Training boxes and the RL reward are noisy, the benchmark has a single annotator and 180 dense images, seven of the eight sources contain public evaluation sets, abstention is evaluated on mostly leaf-centric images, and all fine-tuned models share one backbone family and are evaluated above their training resolution (Appendix R).

AgroGround converts the labels of eight agricultural VQA datasets into a grounding resource larger than any expert-annotated agricultural localization set, together with a human-verified benchmark. Instruction format determines which capabilities fine-tuning preserves, a small share of healthy negatives enables abstention, and a grounded reward improves lesion coverage, with the resulting 2B model exceeding its annotation teacher on our benchmark and on PlantSeg.

Reproducibility Statement

Dataset construction is deterministic given the handler code and fixed seeds (candidate generation 2026, training-data conversion 31, evaluation selection 2024/31/33/44/45 for the annotation rounds, negatives 47 and 51, training seeds 42/1/2, subset sampling 11, pseudo-label audit sample 7, qualitative figure selection 3 for grounding and 11 for recognition, and control experiments 23). Training uses the public qwen-vl-finetune toolchain and verl with the released configurations. Every evaluation number in the paper, in the main text and the appendices alike, is reproducible from the prediction files and scoring script that will be released against the frozen benchmark annotations, whose disjointness from the training split by file path and perceptual hash is verified by the released audit script. Both surviving RL checkpoints are reported.

Ethics Statement

AgroGround is built exclusively from publicly released research datasets, with per-source licensing documented in Appendix Q. Images are referenced by source identifier, not redistributed, and no personal data is collected. Automatic annotations are labeled as machine-generated. Agricultural diagnosis systems should treat model outputs as decision support, not autonomous treatment decisions. The existence-awareness results document both the risk and its mitigation.

LLM Usage Statement

Generative AI tools were used to assist with: feedback on research methodology and experimental design, implementation of data-processing, evaluation, and figure code, drafting and editing parts of the paper to improve readability, polishing certain parts of the paper, identifying and summarizing related literature, brainstorming during the research process, and suggesting structure and keywords. We did not use generative AI tools to generate synthetic data or to formulate mathematical claims or proofs. The research questions, dataset design, and conclusions are the authors' own. All AI-assisted help and text were reviewed by the authors, every reported number was verified against the prediction files and seeded scripts underlying the results, and all statistics were checked.

References

Abderrahmene Boudiaf, Irfan Hussain, and Sajid Javed. AgriChat: A multimodal large language model for agriculture image understanding. *arXiv preprint arXiv:2603.16934*, 2026.

Nicolas Carion, Laura Gustafson, Yuan-Ting Hu, Shoubhik Debnath, Ronghang Hu, Didac Suris, Chaitanya Ryali, Kalyan Vasudev Alwala, Haitham Khedr, Andrew Huang, et al. SAM 3: Segment anything with concepts. In *ICLR*, pp. 138846–138923, 2026. arXiv:2511.16719.

Jinpeng Chen, Runmin Cong, Yuzhi Zhao, Hongzheng Yang, Guangneng Hu, Horace Ho Shing Ip, and Sam Kwong. SEFE: Superficial and essential forgetting eliminator for multimodal continual instruction tuning. In *ICML*, 2025. arXiv:2505.02486.

Jiuqing Dong, Jaehwan Lee, Alvaro Fuentes, Mingle Xu, Sook Yoon, Mun Haeng Lee, and Dong Sun Park. Data-centric annotation analysis for plant disease detection: Strategy, consistency, and performance. *Frontiers in Plant Science*, 13:1037655, 2022. doi: 10.3389/fpls.2022.1037655.

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. In *Advances in Neural Information Processing Systems 38 (NeurIPS 2025), Datasets and Benchmarks Track*, pp. 25023–25099, 2025. doi: 10.52202/085713-0744. arXiv:2506.20100.

Romane Dubois, Lydia Bousset, Stéphane Jumel, Melen Leclerc, Nicolas Parisey, and Alexis Joly. Text guidance is powerful but prompt-sensitive for weakly-supervised leaf symptom segmentation. *Smart Agricultural Technology*, 15:102544, 2026. doi: 10.1016/j.atech.2026.102544. Published version of bioRxiv preprint 10.64898/2026.07.10.737680.

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 38 (NeurIPS 2025), Datasets and Benchmarks Track*, pp. 84185–84217, 2025. doi: 10.52202/085713-2540. arXiv:2504.10568.

Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A. Wichmann. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2:665–673, 2020.

Mohammadreza Haghighat, Alzayat Saleh, and Mostafa Rahimi Azghadi. Multi-label instance-level generalised visual grounding in agriculture. *arXiv preprint arXiv:2603.06699*, 2026.

Shuting He, Henghui Ding, Chang Liu, and Xudong Jiang. GREC: Generalized referring expression comprehension. *arXiv preprint arXiv:2308.16182*, 2023.

Sandesh Hegde, Jaison Saji Chacko, Debarshi Banerjee, and Uma Mahesh. GenSeg-R1: RL-driven vision-language grounding for fine-grained referring segmentation. *arXiv preprint arXiv:2602.09701*, 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 *ICLR*, 2022.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. Segment anything. In *ICCV*, pp. 4015–4026, 2023.

Xin Lai, Zhuotao Tian, Yukang Chen, Yanwei Li, Yuhui Yuan, Shu Liu, and Jiaya Jia. LISA: Reasoning segmentation via large language model. In *CVPR*, 2024.

Haocheng Li, Juepeng Zheng, Zenghao Yang, Kaiqi Du, Guilong Xiao, Gengmeng Pu, Haohuan Fu, and Jianxi Huang. AgroVG: A large-scale multi-source benchmark for agricultural visual grounding. *arXiv preprint arXiv:2605.22034*, 2026a.

He Li, Yuhui Zhang, Xiaohan Wang, Kaifeng Lyu, and Serena Yeung-Levy. Fine-tuning MLLMs without forgetting is easier than you think. *arXiv preprint arXiv:2603.14493*, 2026b.

Jinyang Li, Fengting Zhao, Hongmin Zhao, Guoxiong Zhou, Jiaxin Xu, Mingzhou Gao, Xin Li, Weisi Dai, Honliang Zhou, Yahui Hu, and Mingfang He. A multi-modal open object detection model for tomato leaf diseases with strong generalization performance using PDC-VLD. *Plant Phenomics*, 6:0220, 2024. doi: 10.34133/plantphenomics.0220.

Liunian Harold Li, Pengchuan Zhang, Haotian Zhang, Jianwei Yang, Chunyuan Li, Yiwu Zhong, Lijuan Wang, Lu Yuan, Lei Zhang, Jenq-Neng Hwang, Kai-Wei Chang, and Jianfeng Gao. Grounded language-image pre-training. In *CVPR*, 2022. arXiv:2112.03857.

Qingmei Li, Yang Zhang, Zurong Mai, Yuhang Chen, Shuohong Lou, Henglian Huang, Jiarui Zhang, Zhiwei Zhang, Yibin Wen, Weijia Li, Haohuan Fu, Jianxi Huang, and Juepeng Zheng. Can large multimodal models understand agricultural scenes? Benchmarking with AgroMind. In *Advances in Neural Information Processing Systems 38 (NeurIPS 2025), Datasets and Benchmarks Track*, pp. 175137–175195, 2025. doi: 10.52202/085713-5271. arXiv:2505.12207.

Rang Li, Lei Li, Shuhuai Ren, Hao Tian, Shuhao Gu, Shicheng Li, Zihao Yue, Yudong Wang, Wenhan Ma, Zhe Yang, Jingyuan Ma, Zhifang Sui, and Fuli Luo. GroundingME: Exposing the visual grounding gap in MLLMs through multi-dimensional evaluation. In *CVPR*, pp. 2412–2422, 2026c. arXiv:2512.17495.

Chang Liu, Henghui Ding, and Xudong Jiang. GRES: Generalized referring expression segmentation. In *CVPR*, 2023. arXiv:2306.00968.

Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, and Lei Zhang. Grounding DINO: Marrying DINO with grounded pre-training for open-set object detection. In *ECCV*, pp. 38–55, 2024a. doi: 10.1007/978-3-031-72970-6_3.

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 *ECCV*, pp. 157–170, 2024b. doi: 10.1007/978-3-031-73016-0_10.

Yuqi Liu, Bohao Peng, Zhisheng Zhong, Zihao Yue, Fanbin Lu, Bei Yu, and Jiaya Jia. Seg-Zero: Reasoning-chain guided segmentation via cognitive reinforcement. *arXiv preprint arXiv:2503.06520*, 2025a.

Yuqi Liu, Tianyuan Qu, Zhisheng Zhong, Bohao Peng, Shu Liu, Bei Yu, and Jiaya Jia. VisionReasoner: Unified reasoning-integrated visual perception via reinforcement learning. In *ICLR*, 2026. arXiv:2505.12081.

Ziyu Liu, Zeyi Sun, Yuhang Zang, Xiaoyi Dong, Yuhang Cao, Haodong Duan, Dahua Lin, and Jiaqi Wang. Visual-RFT: Visual reinforcement fine-tuning. In *ICCV*, pp. 2034–2044, 2025b.

Haozhe Luo, Shelley Zixin Shu, Ziyu Zhou, Sebastian Otálora, and Mauricio Reyes. XBench: A comprehensive benchmark for visual-language explanations in chest radiography. In *2026 IEEE 23rd International Symposium on Biomedical Imaging (ISBI)*, pp. 1–5, 2026. doi: 10.1109/isbi61048.2026.11515566. arXiv:2510.19599.

Hazza Mahmood, Yongqiang Yu, and Rao Anwer. AgriChain: Visually-grounded expert-verified reasoning for interpretable agricultural vision–language models. In *Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC)*, pp. 2268–2276. ELRA Language Resource Association, 2026. doi: 10.63317/2bv5k9hnduop. arXiv:2604.07814.

Matthias Minderer, Alexey Gritsenko, and Neil Houlsby. Scaling open-vocabulary object detection. In *NeurIPS*, 2023.

Sharada P. Mohanty, David P. Hughes, and Marcel Salathé. Using deep learning for image-based plant disease detection. *Frontiers in Plant Science*, 7:1419, 2016.

Fuseini Mumuni and Alhassan Mumuni. Segment anything model for automated image data annotation: empirical studies using text prompts from Grounding DINO. *arXiv preprint arXiv:2406.19057*, 2024.

Khang Nguyen Quoc, Phuong D. Dao, and Luyl-Da Quach. LeafNet: A large-scale dataset and comprehensive benchmark for foundational vision-language understanding of plant diseases. *arXiv preprint arXiv:2602.13662*, 2026.

Eric Peh, Debaditya Roy, and Basura Fernando. H-GRPO: Permutation-invariant reinforcement learning for grounded visual reasoning. *arXiv preprint arXiv:2606.29915*, 2026.

Zhiliang Peng, Wenhui Wang, Li Dong, Yaru Hao, Shaohan Huang, Shuming Ma, Qixiang Ye, and Furu Wei. Grounding multimodal large language models to the world. In *ICLR*, 2024.

Qwen Team. Qwen3-VL technical report. *arXiv preprint arXiv:2511.21631*, 2025.

Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, Eric Mintun, Junting Pan, Kalyan Vasudev Alwala, Nicolas Carion, Chao-Yuan Wu, Ross Girshick, Piotr Dollár, and Christoph Feichtenhofer. SAM 2: Segment anything in images and videos. In *ICLR*, 2025.

Syed Nazmus Sakib, Nafiul Haque, Mohammad Zabed Hossain, and Shifat E. Arman. A visual question answering dataset for benchmarking vision-language models in plant science. *Scientific Data*, 13:1161, 2026. doi: 10.1038/s41597-026-07779-y. arXiv:2508.17117 (PlantExpertVQA).

Sraavya Sambara, Sung Eun Kim, Xiaoman Zhang, Luyang Luo, Shreya Johri, Mohammed Baharoon, Du Hyun Ro, and Pranav Rajpurkar. 3DReasonKnee: Advancing grounded reasoning in medical vision language models. In *Pacific Symposium on Biocomputing 2026*, volume 31, pp. 99–113. World Scientific, 2026. doi: 10.1142/9789819824755_0008. arXiv:2510.20967.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. DeepSeekMath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

Haozhan Shen, Peng Liu, Jingcheng Li, Chunxin Fang, Yibo Ma, Jiajia Liao, Qiaoli Shen, Zilun Zhang, Kangjia Zhao, Qianqian Zhang, et al. VLM-R1: A stable and generalizable R1-style large vision-language model. *arXiv preprint arXiv:2504.07615*, 2025.

Risa Shinoda, Nakamasa Inoue, Hirokatsu Kataoka, Masaki Onishi, and Yoshitaka Ushiku. AgroBench: Vision-language model benchmark in agriculture. In *ICCV*, pp. 7634–7644, 2025. doi: 10.1109/ICCV51701.2025.00716.

Davinder Singh, Naman Jain, Pranjali Jain, Pratik Kayal, Sudhakar Kumawat, and Nipun Batra. PlantDoc: A dataset for visual plant disease detection. In *Proceedings of the 7th ACM IKDD CoDS and 25th COMAD*, pp. 249–253, 2020. doi: 10.1145/3371158.3371196.

Wenxuan Wang, Tongtian Yue, Yisi Zhang, Longteng Guo, Xingjian He, Xinlong Wang, and Jing Liu. Unveiling parts beyond objects: Towards finer-granularity referring expression segmentation. In *CVPR*, 2024.

Xuewei Wang and Jun Liu. Enhancing plant disease detection through multi-modal integration of visual and textual data. *Plant Methods*, 22:30, 2026. doi: 10.1186/s13007-026-01521-w.

Tianqi Wei, Zhi Chen, Xin Yu, Scott Chapman, Paul Melloy, and Zi Huang. A large-scale in-the-wild dataset for plant disease segmentation. *Scientific Data*, 13(1):205, 2026. doi: 10.1038/s41597-025-06513-4. Preprint: arXiv:2409.04038 ("PlantSeg: A Large-Scale In-the-wild Dataset for Plant Disease Segmentation").

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. *arXiv preprint arXiv:2511.23253*, 2025.

Linshan Wu, Yuxiang Nie, Sunan He, Jiaxin Zhuang, Luyang Luo, Tao Li, Zhuoyao Xie, Dexuan Chen, Yinghua Zhao, Neeraj Mahboobani, Varut Vardhanabhuti, Ronald Cheong Kin Chan, Yifan Peng, Pranav Rajpurkar, and Hao Chen. A universal foundation model for grounded biomedical image interpretation. *Nature Communications*, 17(1):7173, 2026. doi: 10.1038/s41467-026-73986-1. arXiv:2504.21336.

Size Wu, Sheng Jin, Wenwei Zhang, Lumin Xu, Wentao Liu, Wei Li, and Chen Change Loy. F-LMM: Grounding frozen large multimodal models. In *CVPR*, 2025. arXiv:2406.05821.

Tsung-Han Wu, Giscard Biamby, David Chan, Lisa Dunlap, Ritwik Gupta, Xudong Wang, Joseph E. Gonzalez, and Trevor Darrell. See, say, and segment: Teaching LMMs to overcome false premises. In *CVPR*, pp. 13459–13469, 2024. doi: 10.1109/CVPR52733.2024.01278.

Xiaoping Wu, Chi Zhan, Yu-Kun Lai, Ming-Ming Cheng, and Jufeng Yang. IP102: A large-scale benchmark dataset for insect pest recognition. In *CVPR*, pp. 8787–8796, 2019.

Bin Xiao, Haiping Wu, Weijian Xu, Xiyang Dai, Houdong Hu, Yumao Lu, Michael Zeng, Ce Liu, and Lu Yuan. Florence-2: Advancing a unified representation for a variety of vision tasks. In *CVPR*, pp. 4818–4829, 2024.

Yunfei Xie, Ce Zhou, Lang Gao, Juncheng Wu, Xianhang Li, Hong-Yu Zhou, Sheng Liu, Lei Xing, James Zou, Cihang Xie, and Yuyin Zhou. MedTrinity-25M: A large-scale multimodal dataset with multigranular annotations for medicine. In *ICLR*, 2025. arXiv:2408.02900.

Chen Xin, Andreas Hartel, and Enkelejda Kasneci. DART: An automated end-to-end object detection pipeline with data diversification, open-vocabulary bounding box annotation, pseudo-label review, and model training. *Expert Systems with Applications*, 258:125124, 2024. doi: 10.1016/j.eswa.2024.125124. arXiv:2407.09174.

Mengge Xue, Zhenyu Hu, Liqun Liu, Kuo Liao, Shuang Li, Honglin Han, Meng Zhao, and Chengguo Yin. Strengthened symbol binding makes large language models reliable multiple-choice selectors. In *ACL*, pp. 4331–4344, 2024. doi: 10.18653/v1/2024.acl-long.237. arXiv:2406.01026.

Xuzheng Yang, Jun Ling, Tao Huang, Caiyan Qin, and Peng Wang. Teaching MLLMs to say no: Generalized referring expression comprehension via refusal calibrated GRPO. *arXiv preprint arXiv:2608.04698*, 2026a.

Yang Yang, Huibin Luo, Haotian Wang, Jingchi Jiang, Jie Liu, Jian Wei, and Ming Fang. PestScope: Exclusion-aware large multimodal model for fine-grained agricultural pest segmentation. *IEEE Transactions on Image Processing*, 35:2034–2049, 2026b. doi: 10.1109/TIP.2026.3661417.

Zonglin Yang, Wei-Zhen Liang, Nevin Lawrence, Xin Qiao, Benjamin Riggan, Chi-En Chiang, and Fuchen Li. WeedExpert-R1: Incentivizing botanical reasoning in MLLMs with reinforcement learning for precision weed grounding. *arXiv preprint arXiv:2607.16492*, 2026c.

Haoxuan You, Haotian Zhang, Zhe Gan, Xianzhi Du, Bowen Zhang, Zirui Wang, Liangliang Cao, Shih-Fu Chang, and Yinfei Yang. Ferret: Refer and ground anything anywhere at any granularity. In *ICLR*, 2024.

En Yu, Kangheng Lin, Liang Zhao, Jisheng Yin, Yana Wei, Yuang Peng, Haoran Wei, Jianjian Sun, Chunrui Han, Zheng Ge, Xiangyu Zhang, Daxin Jiang, Jingyu Wang, and Wenbing Tao. Perception-R1: Pioneering perception policy with reinforcement learning. In *NeurIPS*, 2025. arXiv:2504.07954.

Guanqi Zhan, Changye Li, Zhijian Liu, Yao Lu, Yi Wu, Song Han, and Ligeng Zhu. EGM: Efficient visual grounding language models. In *Computer Vision – ECCV 2026*, Lecture Notes in Computer Science, pp. 300–317. Springer, 2026. doi: 10.1007/978-3-032-37044-0_17. arXiv:2601.13633.

Mingqing Zhang, Zhuoning Xu, Peijie Wang, Rongji Li, Liang Wang, Qiang Liu, Jian Xu, Xuyao Zhang, Shu Wu, and Liang Wang. AgriDoctor: A multimodal intelligent assistant for agriculture. In *ICASSP*, pp. 2741–2745, 2026. doi: 10.1109/ICASSP55912.2026.11464537. arXiv:2509.17044.

Chujie Zheng, Hao Zhou, Fandong Meng, Jie Zhou, and Minlie Huang. Large language models are not robust multiple choice selectors. In *ICLR*, 2024. arXiv:2309.03882.

Chujie Zheng, Shixuan Liu, Mingze Li, Xiong-Hui Chen, Bowen Yu, Chang Gao, Kai Dang, Yuqiong Liu, Rui Men, An Yang, Jingren Zhou, and Junyang Lin. Group sequence policy optimization. *arXiv preprint arXiv:2507.18071*, 2025.

Yongshuo Zong, Tingyang Yu, Ruchika Chavhan, Bingchen Zhao, and Timothy Hospedales. Fool your (vision and) language model with embarrassingly simple permutations. In *ICML*, 2024. arXiv:2310.01651.

A CORPUS DETAILS

Dataset	Samples ingested	Groundable	Primary domain
CDDM (Liu et al., 2024b)	1,060,294	513,650	Crop disease QA
LeafNet (Nguyen Quoc et al., 2026)	121,337	81,747	Leaf disease
MIRAGE (Dongre et al., 2025)	40,889	10,560	Expert consultations
AgroMind (Li et al., 2025)	28,482	1,664	Agricultural remote sensing
LeafBench (Nguyen Quoc et al., 2026)	13,950	4,140	Leaf disease VQA
AgroCoT (Wen et al., 2025)	4,535	318	Chain-of-thought reasoning
AgroBench (Shinoda et al., 2025)	4,342	2,810	Disease/pest/weed ID (61%)
AgMMU (Gauba et al., 2025)	772	596	Farmer–expert QA
Total	**1,274,601**	**615,485**	209,740 unique image files

Table 4: Source datasets and cleaning outcome. Groundable and non-groundable (659,116) partitions sum exactly to the ingested total. Sample counts are of the releases we ingested, which differ from the source papers' headline figures in four cases: LeafNet is the public `enalis/LeafNet` training split (121,337 rows) rather than the full 186,000-image dataset. AgroCoT is the v1 release (4,535 samples, then named AgriCoT, with v2 onward reporting 4,759). AgMMU is the released evaluation file as distributed (772 entries, 770 with distinct identifiers) against the 746 questions reported in the paper, and the CDDM figure is our own count of its released question-answer files, the paper stating "137,000 images" and "1 million" question-answer pairs. Two sources carry partial box-level ground truth: AgroBench provides bounding boxes for its 609 weed-identification images to mark which weed each question asks about, and a boundary-analysis subset of AgroCoT embeds box coordinates as text inside answer strings. We use neither as supervision. Every training box comes from the annotation pipeline.

Handlers also repair dataset-specific pathologies: answer templates, abbreviations, and an 87-pattern action/cause vocabulary.

Kind	Samples	Share	Description
Disease	600,905	97.6%	Named diseases with visual symptoms
Object	7,278	1.2%	Identifiable agricultural objects/organisms
Pest	4,103	0.7%	Insect and invertebrate pests
Symptom description	2,543	0.4%	Descriptive symptom queries
Weed	609	0.1%	Weed species
Stress condition	47	$<$0.1%	Abiotic stress
Total	**615,485**	100%	

Table 5: Groundable samples by kind.

Accounting: 1,274,601 ingested $\rightarrow$ 615,485 groundable (209,740 unique files) $\rightarrow$ 614,765 readable $\rightarrow$ 579,260 boxed (188,832 unique files) $\rightarrow$ 794,850 examples. 659,116 non-groundable samples (229,938 unique files) preserved. Train/validation/test unique files: 171,580 / 49,395 / 49,275. Perceptually unique training images: 104,150.

Sources. AgroGround draws on eight agricultural VQA datasets (Figure 1, Table 4). A *sample* is one question–answer pair. Sources ask several questions about the same photograph and, we later found, share photographs with one another, so unique-file and perceptually unique counts are reported alongside sample counts.

Groundability cleaning. Many VQA answers are not visually localizable. Dataset-specific handlers classify every sample by kind (disease, pest, weed, symptom description, stress condition, object, action, cause, healthy) and extract a clean label and crop. Six categories are routed out of the groundable stream: management actions, causes and conditions, aerial field-level labels, CDDM's plant-identification questions (4,115 samples, and species answers from other sources that name a visible organism are retained as object-kind samples), healthy images (154,180 samples), and non-agricultural or over-generic labels. Routed-out samples are preserved with their exclusion reason

(659,116 samples, 229,938 unique files). Healthy images later supply the target-absent evaluation regime and the negative supervision of Section 4.3.

Taxonomy-driven prompts. We do not query the detector with the bare label: it names the disease without the host crop or the visual symptom. We curate a taxonomy of 6,038 disease, pest, and weed labels, each mapped to a descriptive grounding prompt with visual symptom cues. Three corrections proved essential: synonym normalization (23 mappings), crop-aware correction of the taxonomy's default host to the crop actually shown (over 82,000 CDDM and LeafNet samples), and non-leaf overrides for root, branch, and bark diseases. All labels are lowercased consistently wherever they appear, for training instructions and answers, candidate lists, and evaluation prompts, so that no casing pattern distinguishes any label from any other.

Granularity framework. Agricultural grounding operates at two natural granularities: *where the lesions are* (fine) versus *which leaf or plant is affected* (coarse). Conflating them corrupts both training signal and evaluation. Granularity is an established axis in grounding (You et al., 2024; Wang et al., 2024), and the granularity of box annotation measurably changes detector accuracy for plant disease (Dong et al., 2022). Every annotated sample is assigned a level from its box geometry: small localized boxes become `fine_region` examples, a single object-spanning box becomes `coarse_object`, and images supporting both yield one training example at each level with granularity-specific instructions. Within fine-region annotations, boxes covering more than 50% of the image are removed, then class-agnostic NMS at IoU 0.5.

Automatic box annotation. Each groundable sample is annotated by GroundingDINO (Swin-T OGC, open weights) conditioned on its taxonomy prompt (Figure 1), with SAM2 (Hiera-L) converting boxes to masks. Thresholds are box 0.25 / text 0.20, a recall-oriented setting chosen before any evaluation data existed. Neither was tuned on evaluation data, and quality filters remove degenerate boxes (area ratio < 0.001), near-full-image boxes (> 0.95), and empty masks. Multi-instance kinds retain all passing boxes, while single-object kinds keep the highest-confidence box. A box-quality audit found detector confidence too weakly related to box quality to use as a filter (Appendix M), so we filter by geometry (Mumuni & Mumuni, 2024). The same audit found that diffuse diseases legitimately produce whole-leaf boxes, which become coarse-granularity supervision rather than noise. A stratified manual audit of 100 training images (Appendix M) finds every box correct in 62%. Only 9% have boxes on the wrong thing; the rest are granularity or coverage errors, the same tendencies the human-verified evaluation exposes in the detector.

Statistics. Disease dominates the groundable set (97.6%), with pest, weed, symptom, object, and stress kinds providing breadth. The accounting is exact (Figure 1): of 615,485 groundable samples, 614,765 are readable and 579,260 receive at least one valid box (94.2%), spanning 188,832 unique files. Granularity assignment yields 391,758 fine-region and 403,092 coarse-object examples (794,850 total). Because the sources reuse and share photographs, the 171,580 image files of the training split contain 104,150 perceptually unique images (64-bit perceptual hash, Appendix B).

B Near-Duplicate Audit

Every image file is hashed with a 64-bit perceptual hash (pHash, 8×8 DCT). Two images are near-duplicates if their Hamming distance is ≤ 6. Flagged pairs are verified by pixel correlation of 96×96 grayscale thumbnails. Applied to the earlier 1,530-image benchmark against the 171,580 training files: 597 flagged pairs, of which 564 were pixel-verified copies (correlation > 0.97, with 488 at identical pixel size), 8 ambiguous (0.90–0.97, excluded conservatively), and 25 hash collisions (retained). Copies were concentrated in sources that reuse shared pools (LeafNet 209, CDDM 124, LeafBench 121, AgroMind 65). A second pass against the 43,000 healthy negative training images flagged 202 of 387 healthy benchmark images. Round-4 candidate selection rejected 4,824 of 8,661 path-disjoint test-split candidates as near-duplicates of training images, and 7,645 of 13,682 healthy candidates as near-duplicates of training negatives. The final set has zero within-set near-repeats. Perceptual hashing catches copies, resizes, and re-encodings. It does not guarantee absence of heavily cropped or rotated derivatives. Final composition by source: LeafNet 512, CDDM 300, MIRAGE 289, AgroBench 174, LeafBench 122, and 83 from AgroMind, AgMMU, and AgroCoT (43, 35, and 5). By kind, positives span disease (607), pest (188), object or species (163), symptom description (11), weed (3), and stress (3). The 3,463 boxes split 2,790 lesion-level and 673 object-level.

C Annotation Conventions

The disease name and a per-image reference sheet with the taxonomy symptom description were visible during annotation. Boxes are tight to the visible symptom, and all instances are annotated. Dense clusters of merged lesions may be covered by one box over the contiguous affected region, with the region-level metric crediting equivalent partitions. Round-4 additions for field imagery: a named pest that is visible is boxed per insect (fine), a whole plant, tree, or fruit with a systemic condition is one box per affected unit (coarse), and distinct lesions on fruit are boxed individually (fine). A described condition that is not verifiably visible is removed (14 of 452 in round 4), and ambiguous cases are flagged and excluded. Healthy images were visually confirmed in thumbnail review (0 of 320 rejected in round 4b). **Pre-annotation anchoring.** Every round showed GroundingDINO boxes as editable pre-annotations, and each round's manifest preserves them, so anchoring can be measured directly (Table 6). Boxes made an export–import round trip through the annotation tool, which perturbs coordinates by a pixel or two, so a proposal counts as kept at IoU ≥ 0.9 against the final box. The pre-annotations are the annotation pipeline's post-filter boxes, i.e. what the annotator actually saw. Across the 975 positives the detector proposed 1,838 boxes: 66.7% were kept (mean IoU 0.988 over matched pairs), 2.6% were adjusted (IoU 0.5–0.9), and 30.7% were deleted. Of the 3,463 final boxes, 63.2% have no pre-annotation within IoU 0.5 and were drawn by the annotator, and 36.8% are accepted proposals. Round 1 was edited most heavily (42.1% of proposals deleted, 74.0% of its final boxes new). Accepted proposals were rarely tightened, so on that 36.8% the box geometry is the detector's, consistent with the detector's recall against the final set (0.295 in the annotator configuration, 0.411 with taxonomy prompts alone). Twenty round-4 images had been candidates in earlier rounds without being retained. They are counted in round 4, where they were annotated.

Round	Images	Proposals	Kept ($\geq$0.9)	Adjusted	Deleted	Final boxes	Annotator-drawn	IoU (matched)
1	213	356	51.4%	6.5%	42.1%	791	74.0%	0.963
2	106	411	68.6%	0.5%	30.9%	674	57.9%	0.997
3	218	448	72.8%	2.5%	24.8%	791	57.4%	0.991
4	438	623	69.8%	1.9%	28.3%	1,207	63.0%	0.991
All	975	1,838	66.7%	2.6%	30.7%	3,463	63.2%	0.988

Table 6: Pre-annotation anchoring by round: how the annotator treated the detector's editable proposals, and how much of the final ground truth is annotator-drawn.

D Metric Definitions and Input Resolution

Matching and pooling. All (prediction, ground-truth) pairs are sorted by IoU in descending order and consumed greedily, skipping any pair whose prediction or ground truth is already matched and stopping below IoU 0.5. Ties break on $(\text{IoU}, \text{prediction index}, \text{ground-truth index})$ in descending order, which is deterministic but arbitrary. Precision, recall, and F1 are micro-averaged: true positives, false positives, and false negatives are pooled over all positive images and combined once, not averaged per image. Region IoU is the IoU between the unions of predicted and ground-truth boxes, averaged over images. An empty prediction scores 0 and is retained in the average. **Healthy images.** Healthy images do not enter precision, recall, or F1. They are scored only by abstention. Boxes emitted on a healthy image are therefore never counted as false positives, which is generous to models that never abstain. **Recognition.** Recognition and joint accuracy are computed over the 975 positives. Healthy images are excluded from both. On a positive image, an answer of "none", a name absent from the candidate list, or an unparseable answer all count as incorrect, with off-list answers additionally tallied. On a healthy image the name is not scored, only whether the box list is empty. **Input resolution.** The three stages do not share a pixel budget. Supervised fine-tuning caps images at 50,176 pixels ($\approx$0.05 MP, `min_pixels` 784), identically for the 2B, the 8B, and every seed. RL rollout prompts are restricted to images of at most 0.5 MP. This is a selection filter rather than a resize, and it dropped 250 of 19,000 candidate rows. Evaluation passes no cap, so images enter at native resolution. GroundingDINO uses its library transform (shortest side 800, longest 1333) and the two 7B reasoning systems their native 840$\times$840. All models see the same images, so the comparison is fair, but every fine-tuned model is evaluated well above its training resolution.

AgroVG computes a maximum-cardinality bipartite matching and reports macro-aggregated Set-F1 at $\tau \in \{0.50, 0.75\}$, whereas greedy matching can return fewer matches and we report IoU 0.5 only. Appendix E shows the three matching rules agree to three decimals on every row.

Validation split. The corpus has a sample-level validation split (79,464 examples), which was not used for model selection: supervised runs take the final checkpoint with no checkpoint selection (one epoch for the main models, three for the small-subset baselines of Appendix N), and the RL checkpoint rule used held-out training-split prompts scored against pseudo-labels (Appendix J). Because the split is made at sample level it shares photographs with the training partition (the three partitions' 270,250 file slots span only 188,832 distinct files), so it measures fit rather than generalization.

Candidate normalization. After normalization, a candidate is rejected if it equals another or is a substring of another, so "leaf spot" cannot accompany "bacterial leaf spot" while distinct diseases sharing words are kept.

Candidate protocol. Given an image and a candidate set, the model must identify which disease is present and output boxes supporting the diagnosis, or an empty list if none apply (the closed-set formulation of biomedical grounded interpretation, Section 2). Each benchmark image receives four lowercased candidates: the true label plus three distractors drawn from the same crop's diseases in the corpus (dataset- and kind-level fallbacks where the crop has fewer than three), shuffled per image with a fixed seed, with a substring rule preventing overlapping candidates. Healthy images receive four plausible candidates with correct answer "none". Because 36% of benchmark positives name a pest, weed, or identifiable organism rather than a disease (Section 3.3), we use "disease" as shorthand for the named condition throughout. We report recognition accuracy (chance 25%), the position distribution of the model's picks (a probe for prompt-order bias), and joint accuracy: correct disease *and* region IoU ≥ 0.5.

E Metric Robustness and Additional Protocols

Placement of the gains. The margin comes from instance recovery rather than box tightness: at IoU 0.75 the taxonomy-prompt detector leads (0.375 vs. 0.342). The gains are unevenly placed: on the 180 dense images the RL model leads every baseline (F1 0.367 vs. 0.160 for the zero-shot 8B). On the 795 non-dense images the zero-shot 8B is best (0.643 vs. 0.603–0.629).

The matching rule is immaterial: greedy, optimal-assignment (Hungarian), and maximum-cardinality matching give identical F1@0.5 to three decimals for every row of Table 8. Table 7 rescores the disease-given protocol under alternative metric choices. Macro averaging, which weights every image equally, compresses the gaps.

Model	Micro F1	Macro F1	Dense F1	Non-dense F1	F1@0.75
GroundingDINO (raw)	0.401	0.577	0.253	0.553	0.311
GroundingDINO (taxonomy)	0.395	0.475	0.315	0.459	**0.375**
VisionReasoner-7B	0.369	0.595	0.219	0.542	0.224
Seg-Zero-7B	0.247	0.504	0.012	0.519	0.206
Qwen3-VL-2B (zero-shot)	0.320	0.584	0.076	0.602	0.239
Qwen3-VL-8B (zero-shot)	0.380	0.615	0.160	**0.643**	0.270
AgroGround-2B, grounding-only	0.426	0.635	0.258	0.629	0.309
AgroGround-2B, +recognition	0.435	0.635	0.291	0.617	0.308
AgroGround-2B, +existence 12%	0.440	0.612	0.316	0.603	0.288
AgroGround-2B, +GRPO	**0.477**	**0.647**	**0.367**	0.619	0.342
AgroGround-8B, +existence 12%	0.429	0.619	0.285	0.610	0.317

Table 7: Metric variants, disease-given protocol. Dense = the 180 benchmark images with $\geq$5 boxes. Macro F1 averages per image over the 975 positives.

Scored on candidate prompts with predicted labels ignored, localization survives grounding-only training: grounding-only F1 0.441 (region IoU 0.606), +recognition 0.432 (0.609), +existence 0.438 (0.595), +GRPO 0.468 (0.606). Under stricter joint criteria, instance-level joint (correct name and at least one matched box) gives 24.4 / 58.7 / 59.9 / 55.8 / 63.4% for the grounding-only, +recognition, +existence, RL, and 8B models, preserving the ordering of Table 9, and fine-only joint accuracy

on the 508 lesion-level images gives 15.9 / 29.5 / 30.9 / 32.1 / 35.6%, where the RL model moves ahead of the supervised 2B models, consistent with its lesion-level grounding gains. The region-IoU criterion in the main text is the strictest of the three and remains the headline metric.

F Recognition and Abstention Controls

Blank and shuffled images. Re-running the candidate protocol on the 975 positives with the image replaced by uniform gray separates what recognition uses. The +recognition model scores 34.1% [31.0, 37.2] against a 25% chance level: a learned prior of about nine points from candidate construction, with image content supplying the remaining 38.5 points of its 72.6%. The existence-trained models mostly abstain on blank images (85.5% and 87.9% empty output for the +existence and RL models, scoring 5.6% and 4.7% with empties counted as errors), treating a contentless image as target-absent. The zero-shot 2B scores 16.3% [14.2, 18.8] with 31.7% empty output, i.e. chance among answered items: the base model carries no prior over our candidate lists. With the image present but tile-shuffled on a 4×4 grid, accuracy is largely preserved (69.7 / 71.2 / 64.9% against 72.6 / 74.7 / 69.2% on real images), so recognition relies on local lesion texture rather than global leaf structure. **Hard negatives.** Queried for a plausible same-crop disease that is *not* present on a diseased image, where the correct output is an empty list, the +existence model abstains on 15.2% [12.9, 17.2] and the RL model on 7.2% [5.5, 8.9] of images. The +recognition model, trained without negatives, abstains on 0.3%. Against their 95.0% and 87.7% abstention on healthy images, this shows the acquired behavior is health detection rather than verification of the named target. Negatives pairing diseased images with absent-disease queries are the direct remedy and are future work.

G Baseline Adaptations

VisionReasoner-7B and Seg-Zero-7B are run with their published prompt templates, 840×840 resize, and native box outputs (masks are a SAM2 refinement of those boxes, so boxes are the fair comparison), scaled back to original pixels. For the candidate protocol, which these grounding models cannot attempt natively, the candidate first mentioned in their reasoning text is taken as their answer: VisionReasoner-7B scores 40.6% [37.6, 43.7], joint 26.8%, with a 61.0% first-position share indicating that the heuristic often captures a restatement of the list rather than a decision; Table 2 marks these values with a dagger. GDINO's candidate-protocol adaptation runs one pass per candidate and answers with the highest-confidence candidate (27.6%).

H Full Results

Rows use the configuration names from development: "+recognition" is the mixed configuration and "+existence" is mixed + healthy (12%). The 6% and 12% mixtures total 678,977 and 721,977 examples.

I Per-Source Results

Teacher configurations. The teacher is scored with the same raw disease name the VLM receives (0.401 [0.368, 0.434]), with the annotation-time taxonomy prompts that produced the training labels (0.395 [0.374, 0.416]), and with those prompts plus the annotation pipeline's geometry filters (0.392 [0.363, 0.425]). With taxonomy prompts alone it reaches the highest recall of any system (0.411) at precision 0.380, while every fine-tuned model trades recall for precision (0.70–0.76). With raw prompts it locks onto the salient diseased leaf and blankets it, adequate when the leaf is the target (coarse F1 0.608). Taxonomy prompts fragment it into more, less precise boxes, gaining on leaf-image sources and losing on whole-plant field sources. The geometry filters restore precision (0.584) and whole-object coverage (coarse F1 0.700) but cut recall to 0.295.

J Reward Function and RL Details

SFT configuration. LoRA rank 64, α=128 on attention and MLP projections, lr 10^{-4}, cosine schedule, effective batch 32, one epoch, taking ∼5.5 h on 2×A6000.

Model	F1@0.5 [95% CI]	R	P	RIoU	Fine F1	Coarse F1	Abst.	False abst.
GroundingDINO (raw prompts)	0.401 [0.368, 0.434]	0.333	0.503	0.594	0.323	0.608	0.0%	0.1%
GroundingDINO (annotation-time taxonomy prompts)	0.395 [0.374, 0.416]	**0.411**	0.380	0.537	0.358	0.476	0.0%	0.0%
GroundingDINO + annotation-pipeline filters (annotator config)	0.392 [0.363, 0.425]	0.295	0.584	0.586	0.282	**0.700**	20.2%	0.5%
VisionReasoner-7B (Liu et al., 2026)	0.369 [0.340, 0.401]	0.314	0.448	0.618	0.302	0.563	0.2%	0.5%
Seg-Zero-7B (Liu et al., 2025a)	0.247 [0.223, 0.276]	0.159	0.563	0.557	0.124	0.604	0.0%	0.0%
Qwen3-VL-2B (zero-shot)	0.320 [0.288, 0.357]	0.220	0.584	0.614	0.205	0.661	1.2%	0.0%
Qwen3-VL-8B (zero-shot)	0.380 [0.347, 0.418]	0.269	0.645	**0.644**	0.294	0.651	30.9%	2.5%
AgroGround-2B, grounding-only	0.426 [0.394, 0.459]	0.296	**0.760**	0.622	0.366	0.619	0.0%	0.0%
AgroGround-2B, +recognition (§4.2)	0.435 [0.404, 0.469]	0.311	0.727	0.617	0.377	0.630	0.2%	0.0%
AgroGround-2B, +existence 12% (§4.3)	0.440 [0.409, 0.474]	0.320	0.708	0.578	0.392	0.614	95.0%	4.6%
AgroGround-2B, +GRPO (§4.4)	**0.477** [0.448, 0.509]	0.362	0.700	0.618	**0.439**	0.617	87.7%	1.0%
AgroGround-8B, +existence 12%	0.429 [0.401, 0.461]	0.301	0.747	0.596	0.369	0.631	**96.0%**	2.9%

Table 8: Grounding with the disease given, on 1,480 human-verified images (975 positives, 505 healthy). RIoU = region-level union IoU. Abst. / false abst. = empty output on healthy / diseased images. Existence models' abstention uses a crop-matched disease query per healthy image. The AgroGround-2B rows are the seed-42 run, the best of three seeds on F1, recognition, joint accuracy, and abstention (Appendix N). GroundingDINO is scored with raw names, with the taxonomy prompts that produced the training labels, and with the annotation pipeline's filters.

Model	Recognition [95% CI]	Pos.-1 share	Joint [95% CI]	Abst.	False abst.
GroundingDINO (best candidate)	27.6% [24.8, 30.4]	26.6%	17.5% [15.4, 19.9]	0.0%	0.0%
Qwen3-VL-2B (zero-shot)	51.8% [48.4, 54.8]	33.7%	34.5% [31.5, 37.4]	0.0%	3.0%
Qwen3-VL-8B (zero-shot)	64.4% [61.1, 67.1]	26.4%	44.6% [41.5, 47.9]	31.1%	2.6%
AgroGround-2B, grounding-only	29.1% [26.3, 32.1]	88.8%	19.2% [16.8, 21.7]	0.0%	1.8%
AgroGround-2B, +recognition format	72.6% [70.0, 75.4]	26.7%	43.3% [40.2, 46.4]	0.0%	0.0%
AgroGround-2B, +existence 12%	74.7% [71.9, 77.5]	25.1%	44.8% [41.7, 48.0]	89.1%	0.9%
AgroGround-2B, +GRPO	69.2% [66.4, 72.0]	24.5%	43.1% [40.0, 46.1]	88.5%	1.2%
AgroGround-8B, +existence 12%	**78.2%** [75.4, 80.7]	25.7%	**48.6%** [45.7, 51.6]	**94.7%**	2.8%

Table 9: Grounded disease recognition on 1,480 images (975 positives, 505 healthy), candidate protocol. Pos.-1 = share of in-list picks that are the first-listed candidate (uniform = 25%). Joint = correct disease and region IoU ≥ 0.5. VisionReasoner's adapted score is reported in Table 2 and Appendix G.

Two residual failures remain after supervised training, both reward-shaped rather than architecture-limited: lesion under-enumeration (the best supervised model emits 3.6 boxes on dense images that contain 12.2 lesions on average, set recall 0.320) and residual false abstention (4.6%). We post-train the mixed + healthy (12%) model with GRPO (Shao et al., 2024) (Algorithm 1), adapting the Qwen3-VL grounding recipe of Zhan et al. (2026) with 8 rollouts per prompt (hyperparameters below), on 18,750 training-split prompts that oversample dense multi-lesion images (50% of positives) and include 16% healthy images, with pseudo-labels as reward targets and every benchmark image excluded. The composite reward has five terms, a format gate, localization, multi-instance recall, recognition, and two-sided abstention (Algorithm 1). We train two epochs (292 steps on 2×A6000). Two checkpoints of this run exist, step 170 (merged mid-run) and the final step 292. Validation reward on held-out training-split prompts favored step 170 (0.867 vs. 0.859); both were evaluated on the benchmark and both are reported (Appendix K), with step 170 as the main result and, since no step-146 checkpoint was retained, not exactly like-for-like with the one-epoch ablations.

Reward $= \mathbf{1}[\text{format}] \cdot \big(0.10 + 0.35\,\ell + 0.15\,m + 0.25\,r + 0.15\,a\big)$, where ℓ is the mean of instance-level F1 (greedy IoU$\geq$0.5 matching) and region-union IoU against the pseudo-label boxes, m is instance recall, r is 1 for the correct candidate, 0.25 for a listed-but-wrong candidate, 0 otherwise (on disease-given prompts its weight is redistributed to ℓ and m), and a is 1 iff the output is empty on healthy inputs and non-empty on diseased inputs. Healthy examples are scored by the gate and a only. Coordinates are clipped to $[0, 1000]$, and boxes below 0.1% of the image count as false positives. Configuration adapts Zhan et al. (2026): GRPO with the GSPO loss variant, lr 3×10^{-6} with 10 warmup steps, low-variance KL (0.005), 8 rollouts per prompt at temperature 0.7 (top-p 0.9, top-k 50), batch 128 prompts, prompt/response budgets 1024/512 tokens, and rollout prompts restricted to images of at most 0.5 MP, a selection filter rather than a resize. Rollout data: 18,750 prompts, disjoint from the benchmark, comprising 3,952 candidate-format and 3,950 disease-given positives from regular images, 3,960 and 3,941 from dense ($\geq$4 pseudo-label boxes) images, and 1,472 and 1,475 healthy negatives. The intended quotas were 4,000 per positive cell and 1,500 per

Model	LeafNet (342)	MIRAGE (289)	AgroBench (174)	CDDM (80)	AgMMU (35)	AgroMind (33)	LeafBench (20)
GroundingDINO	20.2 / 0.394	32.9 / 0.585	31.0 / 0.270	33.8 / 0.339	31.4 / 0.333	27.3 / 0.598	20.0 / 0.278
GroundingDINO (taxonomy prompts)	– / 0.402	– / 0.453	– / 0.327	– / 0.392	– / 0.345	– / 0.441	– / 0.329
Qwen3-VL-8B (zero-shot)	53.8 / 0.352	**84.4** / **0.598**	**65.5** / 0.247	52.5 / 0.418	48.6 / 0.362	57.6 / 0.396	35.0 / 0.341
AgroGround-2B, +recognition	85.1 / 0.439	69.6 / 0.593	50.0 / 0.293	98.8 / 0.501	45.7 / 0.365	72.7 / 0.477	50.0 / 0.356
AgroGround-2B, +existence 12%	85.4 / 0.443	74.7 / 0.593	48.9 / 0.311	**100.0** / 0.495	**62.9** / 0.252	66.7 / 0.471	55.0 / 0.458
AgroGround-2B, +GRPO	80.4 / **0.497**	65.1 / 0.570	48.3 / **0.344**	95.0 / **0.539**	51.4 / **0.369**	66.7 / **0.552**	60.0 / **0.580**
AgroGround-8B, +existence 12%	**86.5** / –	83.7 / –	48.3 / –	98.8 / –	**62.9** / –	**75.8** / –	**65.0** / –

Table 10: Per-source recognition accuracy (%, candidate protocol) / disease-given F1@0.5 on the benchmark positives (number of positives in parentheses, AgroCoT's 2 positives omitted). The fine-tuned 2B models dominate on in-domain leaf sources, while the zero-shot 8B base is stronger at recognition on field sources. The fine-tuned 8B combines both on MIRAGE but not on AgroBench (recognition only shown for it, grounding only for the taxonomy-prompt GroundingDINO).

Algorithm 1 AgroGround training

1: **Input:** pseudo-labeled corpus $\mathcal{D}^+$ (794,850 examples, 635,977 in the training split), healthy image pool $\mathcal{H}$, base model θ_0 (Qwen3-VL-2B or -8B), rollout prompt set $\mathcal{R}$ (18,750 training-split prompts: dense-oversampled positives, 16% healthy)
2: **Stage 1: supervised fine-tuning**
3: Build the mixture $\mathcal{M}$: each training example is rewritten in the candidate format with probability 0.5 (four crop-matched candidates, correct one at a random position) and otherwise keeps the disease-given instruction; add 86,000 negatives from $\mathcal{H}$ (43,000 images $\times$ two formats, answer `[]`), a 12% share
4: Train LoRA adapters (r=64, α=128) on the language model with next-token cross-entropy for one epoch $\rightarrow \theta_{\mathrm{SFT}}$
5: **Stage 2: reinforcement learning with a composite grounded reward** (2B model)
6: **for** each batch of prompts $x \sim \mathcal{R}$ **do**
7: Sample G=8 responses $y_1, \ldots, y_G \sim \pi_\theta(\cdot \mid x)$
8: Score each response against the pseudo-labels: $R(y) = \mathbf{1}[\text{valid format}]\,\big(0.10 + 0.35\,\ell + 0.15\,m + 0.25\,r + 0.15\,a\big)$; on disease-given prompts r's weight is redistributed to ℓ and m, and healthy prompts are scored by the format gate and a only
9: Compute group-relative advantages and update θ with GRPO using the GSPO loss variant and a KL penalty (β=0.005) toward θ_{SFT}
10: **end for**
11: **Output:** of the two retained checkpoints (step 170, merged mid-run, and the final step 292), the one with the higher validation reward on held-out training-split prompts (step 170; both are reported)

negative cell, and 250 rows in total were removed by the 0.5 MP filter. Checkpoints were written every 10 steps with only the two most recent retained. Step 170 was additionally merged mid-run. Validation reward: 0.845 $\rightarrow$ 0.879 at step 30 $\rightarrow$ 0.867 at step 170 $\rightarrow$ 0.859 at step 292. Steps 170 and 292 were both evaluated on the benchmark (Appendix K).

K RL Ablations and Checkpoint Consistency

Configuration	F1@0.5 [CI]	R	Fine F1	Boxes (dense)	Abst. given / cand.	False abst.	Recog.	Joint
SFT initialization (+existence 12%)	0.440 [0.409, 0.474]	0.320	0.392	3.62	**95.0** / 89.1	4.6%	**74.7%**	**44.8%**
Main reward, dense-oversampled rollouts, step 170 (reported)	0.477 [0.448, 0.509]	0.362	0.439	4.27	87.7 / 88.5	**1.0%**	69.2%	43.1%
Main reward, final checkpoint (2 epochs, step 292)	0.467 [0.436, 0.497]	0.347	**0.440**	4.02	89.3 / 95.2	6.5%	67.2%	39.0%
Recognition-weighted reward (0.30/0.10/0.35), 1 epoch	**0.480** [0.452, 0.514]	**0.367**	0.436	**4.41**	84.2 / 87.1	3.0%	70.2%	44.1%
Main reward, uniform rollout mix, 1 epoch	0.440 [0.413, 0.473]	0.319	0.409	3.58	88.1 / 90.1	4.9%	70.2%	40.9%

Table 11: RL configurations on the benchmark. Validation-reward curves: main run 0.845$\rightarrow$0.879 (30)$\rightarrow$0.867 (170)$\rightarrow$0.859 (292), recognition-weighted 0.858$\rightarrow$0.885 (30)$\rightarrow$0.883 (146), uniform-mix 0.853 (146). Both surviving checkpoints of the main run (170 and 292) are shown. Dense oversampling drives the localization gain, the recognition cost appears in every configuration, and the final checkpoint is lower on F1, recall, recognition, and joint accuracy and higher on false abstention.

L Open-Set Naming

Without a candidate list, the model is asked to name the specific disease or pest and box the evidence. Under a first prompt ("identify the disease, pest, or condition"), models copied the category words as labels. Under a second prompt giving example names, they copied the examples. Lenient match rate (predicted name contains or is contained in the ground-truth label): zero-shot 2B 4.6%, grounding-only 2.5%, +recognition 2.7%, +existence 2.5%, RL 2.4%. Strict match is $\leq 2.1\%$ for all. Grounding F1 without the name: 0.22 (zero-shot) to 0.31–0.33 (RL models). The existence-aware models abstain on 83–93% of healthy images but also on 6–9% of diseased ones. Unprompted naming is not a capability of these models. The acquired capability is closed-set discrimination.

M Pseudo-Label Audit

100 pseudo-labeled training images were sampled by stratifying the box-annotated corpus over the eight largest source–granularity strata (seed 7): CDDM coarse 13, CDDM fine 13, LeafNet fine 13, LeafNet coarse 13, MIRAGE fine 13, LeafBench coarse 13, AgroBench fine 11, LeafBench fine 11, and judged with the detector's boxes overlaid, under the same conventions as the benchmark annotation. Verdicts (no audited image fell into more than one issue category): all boxes correct 62, whole-leaf box on a lesion-level image (too large) 9, some lesions unboxed (missed) 10, a whole-leaf box duplicating a region box 10, boxes on something other than the labeled condition (wrong) 9, and images with any issue 38. Detector confidence is only weakly related to box quality: mean box confidence is 0.425 on the 62 images where every box is correct and 0.372 on the 38 flagged images (Spearman $\rho = 0.17$, $p = 0.08$). Using the maximum box confidence instead gives $\rho = 0.05$, $p = 0.65$. We therefore filter by geometry rather than by confidence. The sampled list and verdicts are released with the audit script.

N Seed Variance, Scaling, Expert-Label Baseline, and 8B Model

Expert labels, pseudo-labels, and scale. Are 794,850 pseudo-labeled examples worth more than 11K expert masks? We fine-tune the same 2B model with the same grounding-only recipe on PlantSeg's expert masks converted to boxes (9,120 examples after removing near-duplicates of the benchmark) and on an equal-size random subset of AgroGround. On the benchmark the two are indistinguishable (F1 0.376 vs. 0.372, paired difference -0.004, 95% CI [-0.024, +0.017]), with the expert boxes giving better recall and region IoU and the pseudo-boxes higher precision, even though PlantSeg is out-of-domain for this test set. Scale then adds what expert labels at that size cannot: the full corpus improves on the equal-size subset by +0.054 F1 (CI [+0.003, +0.098]) and on a 10K one-epoch run by +0.026 (CI [+0.008, +0.045]), with the gain concentrated at lesion level (fine F1 $0.27 \rightarrow 0.37$). Between 10K and 100K examples F1 does not increase (0.400 to 0.388). The equal-size runs use three epochs and the larger ones one, because small sets were given more passes rather than matched compute, so data scale and optimization budget are not separated, and all subset runs are single-seed. **Model scale.** The same existence recipe on Qwen3-VL-8B gives F1 0.429 (paired difference vs. the 2B -0.011, not significant) with higher precision and *fewer* boxes per image (1.88 vs. 2.23): scale does not buy lesion enumeration, and Table 9 shows what it does buy. **External transfer.** On PlantSeg's test set (2,165 images after removing 119 near-duplicates of our training files, with expert masks converted to boxes), every AgroGround model transfers on instance F1, though not on region IoU, where the AgroGround models (0.367–0.463) fall below the zero-shot bases (0.497 and 0.547). The RL model (F1 0.282) outperforms GDINO (0.182), both zero-shot bases (0.164, 0.215), and a model trained on PlantSeg's own expert masks (0.251, paired +0.031 [+0.018, +0.042]). The RL gain over its initialization replicates out of domain (+0.051, CI [+0.042, +0.059], Appendix O).

Seeds. Three seeds (42, 1, 2) of the +existence 12% model on the benchmark: F1 0.440 / 0.436 / 0.432 (mean 0.436 ± 0.004), recall 0.316 ± 0.003, disease-given abstention $94.1 \pm 0.9\%$, recognition 74.7 / 73.8 / 72.4 (73.6 ± 1.1), and joint 43.5 ± 1.2. Every effect reported in the main text exceeds this variance by a wide margin. The 6%-vs-12% negative-share difference does not.

8B model. Qwen3-VL-8B fine-tuned with the +existence 12% recipe (identical data and hyperparameters, per-device batch 2 with 8 accumulation steps for the same effective batch of 32): disease-

Negatives	Abst. (given) [CI]	Abst. (cand.)	False abst. (given / cand.)	R	F1	RIoU	Recog.	Joint
0%	0.2% [0.0, 0.6]	0.0%	0.0% / 0.0%	0.311	0.435	0.617	72.6%	43.3%
6%	93.3% [90.7, 95.5]	89.5%	5.2% / 0.5%	0.310	0.433	0.585	73.0%	44.3%
12%	**95.0%** [93.1, 96.9]	89.1%	4.6% / 0.9%	**0.320**	**0.440**	0.578	**74.7%**	**44.8%**

Table 12: Existence-aware supervision (505 healthy images). Negatives buy near-complete abstention at a small false-abstention cost. Differences between 6% and 12% are within seed variance.

Training data (grounding-only recipe)	F1@0.5 [CI]	R	P	RIoU	Fine F1	Coarse F1
PlantSeg expert boxes, 9,120 ex., 3 ep.	0.376 [0.345, 0.407]	0.275	0.595	**0.688**	0.282	0.691
AgroGround pseudo-boxes, 9,120 ex., 3 ep.	0.372 [0.340, 0.409]	0.249	0.734	0.639	0.270	0.682
AgroGround pseudo-boxes, 10K ex., 1 ep.	0.400 [0.367, 0.439]	0.275	0.734	0.647	0.306	**0.698**
AgroGround pseudo-boxes, 100K ex., 1 ep.	0.388 [0.358, 0.424]	0.258	**0.784**	0.635	0.299	0.660
AgroGround pseudo-boxes, 635,977 ex., 1 ep.	**0.426** [0.394, 0.459]	**0.296**	0.760	0.622	**0.366**	0.619

Table 13: Expert labels vs. pseudo-labels at equal size, and scaling, on the benchmark (disease-given protocol). Paired differences: pseudo minus expert at equal size −0.004 [−0.024, +0.017], full corpus minus equal-size subset +0.054 [+0.003, +0.098]. Subset runs are single-seed.

given F1 0.429 [0.401, 0.461], R 0.301, P 0.747, RIoU 0.596, abstention 96.0% (false abstention 2.9%), 1.88 boxes per lesion image. Candidate protocol: recognition 78.2% [75.4, 80.7], joint 48.6%, abstention 94.7%, false abstention 2.8%. Paired vs. the 2B: F1 −0.011 [−0.050, +0.027], recognition +3.5 points [+0.7, +6.1].

O External Transfer to PlantSeg

We use the Zenodo release of 15 September 2024 (record 10.5281/zenodo.13762907, 11,458 images over 115 disease categories, CC BY-NC-ND 4.0). The journal version of the paper reports 7,774 diseased images with masks and states that roughly 15% of annotated images undergo expert review, with disputed images adjudicated by plant pathologists.

Model	F1@0.5 [CI]	R	P	RIoU	Fine F1	Coarse F1
GroundingDINO (raw prompts)	0.182 [0.172, 0.194]	0.140	0.260	0.456	0.147	0.617
GroundingDINO (taxonomy prompts)	0.183 [0.173, 0.194]	0.144	0.250	0.446	0.149	0.610
GroundingDINO + annotation-pipeline filters (annotator config)	0.161 [0.150, 0.173]	0.104	0.351	0.371	0.133	0.655
Qwen3-VL-2B (zero-shot)	0.164 [0.152, 0.177]	0.104	0.389	0.497	0.118	0.892
Qwen3-VL-8B (zero-shot)	0.215 [0.201, 0.231]	0.140	0.463	0.547	0.173	0.899
AgroGround-2B, grounding-only	0.231 [0.219, 0.245]	0.153	**0.480**	0.423	0.189	0.940
AgroGround-2B, +existence 12%	0.231 [0.219, 0.245]	0.161	0.412	0.377	0.197	0.871
AgroGround-2B, +GRPO	**0.282** [0.269, 0.296]	**0.213**	0.418	0.463	**0.249**	0.935
AgroGround-8B, +existence 12%	0.237 [0.224, 0.251]	0.159	0.468	0.367	0.203	0.863
PlantSeg-trained 2B (in-domain reference)	0.251 [0.238, 0.266]	0.178	0.426	**0.588**	0.211	**0.981**
AgroGround pseudo-boxes, 9,120 ex. (equal size)	0.188 [0.176, 0.201]	0.118	0.459	0.451	0.142	0.920

Table 14: PlantSeg test set, 2,165 images (1,750 lesion-level, 415 object-level), disease-given protocol with PlantSeg's disease names. The released test split contains 2,295 images. 8 have empty masks and 3 have no connected component covering at least 0.1% of the image, leaving 2,284 usable, of which 119 are perceptual-hash near-duplicates of our training files and are removed. Masks are matched to images by file name and converted to component boxes. Taxonomy prompts were used for the 363 test images whose PlantSeg disease name maps to a taxonomy entry. The remaining 1,802 fall back to the raw name, so the taxonomy-based rows differ little from the raw-prompt row. Paired differences: RL minus its initialization +0.051 [+0.042, +0.059], RL minus the annotator configuration +0.121 [+0.109, +0.133], RL minus the PlantSeg-trained model +0.031 [+0.018, +0.042], RL minus the taxonomy-prompt teacher +0.099 [+0.089, +0.110], 8B minus 2B +0.006 [−0.004, +0.015], and PlantSeg-trained minus equal-size pseudo +0.063 [+0.053, +0.075]. Object-level F1 is near-trivial here because many masks cover the whole plant. The grounding-only and +existence models tie at F1 0.231 with the same bootstrap interval. The tie is genuine, and they differ in recall and precision (0.153/0.480 vs. 0.161/0.412).

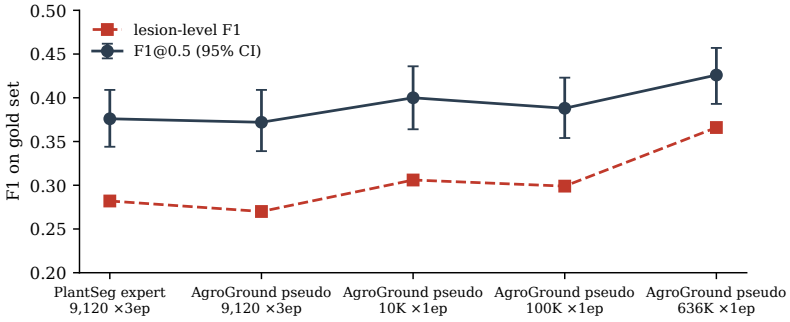

Figure 5: The same data as the table above: F1@0.5 with 95% CIs and lesion-level F1 against training data. Scale helps mainly at lesion level.

P QUALITATIVE EXAMPLES

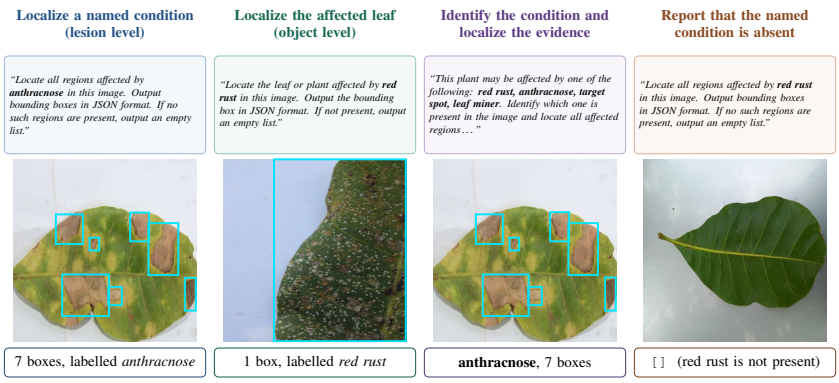

Figure 6: The tasks AgroGround trains and evaluates, one column per task, shown with the outputs of the main supervised model (AgroGround-2B, mixed + healthy) on benchmark cashew leaves. Boxes are the model's predictions, and prompts are the evaluation prompts, abbreviated in the third panel.

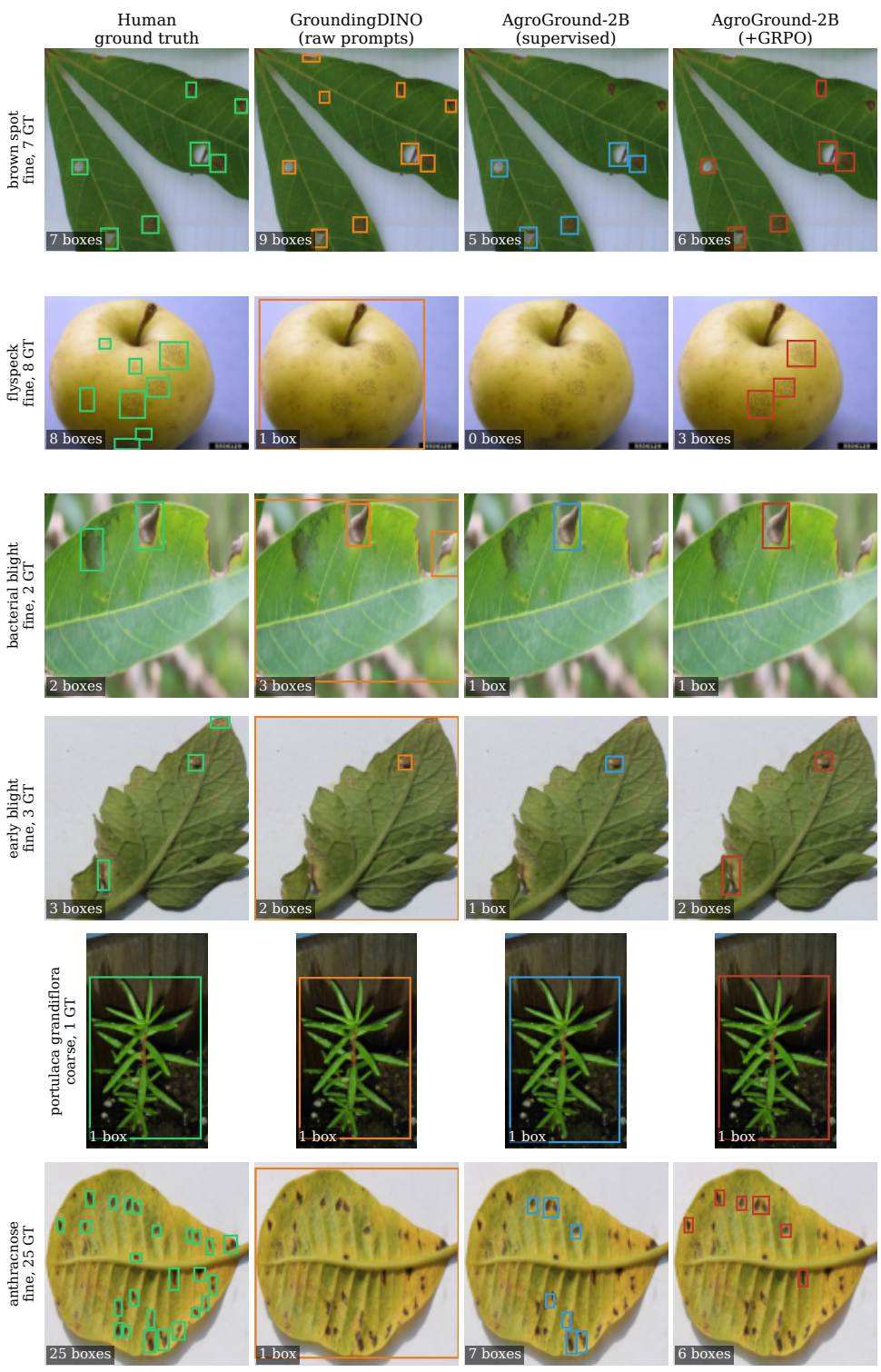

Figure 7: Disease-given grounding on six benchmark images selected with a fixed seed (3) under stated rules: two dense lesion-level images where the RL model's region IoU exceeds 0.5, two with 2–4 lesions, one object-level image, and one dense failure (at least 8 lesions, RL region IoU below 0.35). Columns: human ground truth (green), GroundingDINO with raw prompts (orange), the supervised AgroGround-2B (blue), and AgroGround-2B after GRPO (red). Numbers are box counts. The detector over-covers with leaf-spanning boxes (the flyspeck, early-blight, and anthracnose rows). The supervised model emits no box on the flyspeck apple, where RL recovers three, and on the 25-lesion anthracnose leaf both fine-tuned models under-enumerate (7 and 6 boxes).

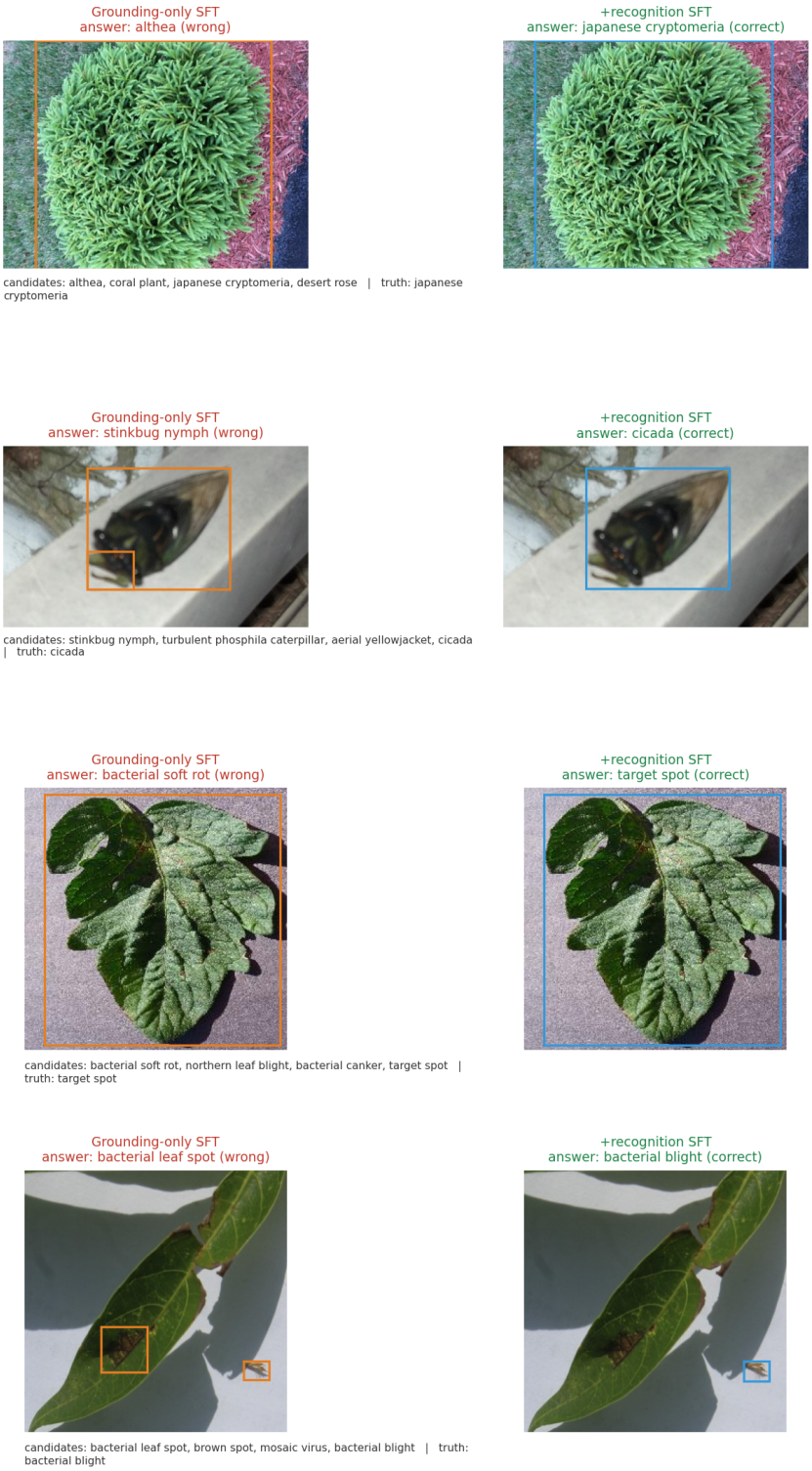

Figure 8: Grounded disease recognition on benchmark images, selected with a fixed seed from cases where the +recognition model is correct, the grounding-only model is wrong, and the grounding-only answer is the first-listed candidate. The grounding-only model echoes that first candidate while localizing plausibly. The +recognition model names the target from the image. Two of the four examples name a species or a pest rather than a disease, reflecting the composition of the benchmark positives (Section 3.3).

Q LICENSES AND RELEASE

Source	Distribution	License (as stated by the source)	Our use
CDDM (Liu et al., 2024b)	GitHub (UnicomAI/UnicomBenchmark)	None stated[a]	training, evaluation
LeafNet / LeafBench (Nguyen Quoc et al., 2026)	Hugging Face (gated)	Tagged CC BY 4.0 / MIT; non-commercial gate[b]	academic research use under the gate terms
MIRAGE (Dongre et al., 2025)	GitHub, Hugging Face	CC BY-SA 4.0	training, evaluation
AgroMind (Li et al., 2025)	GitHub, Hugging Face	None stated[c]	training, evaluation
AgroCoT (Wen et al., 2025)	Hugging Face	CC BY-SA 4.0 (dataset-card metadata)[d]	training, evaluation
AgroBench (Shinoda et al., 2025)	Hugging Face (gated)	None stated[e]	training, evaluation
AgMMU (Gauba et al., 2025)	Hugging Face	CC BY-SA 4.0 in the paper; data tagged `cc`[f]	training, evaluation
PlantSeg (Wei et al., 2026)	Zenodo (DOI 10.5281/zenodo.13762907)	CC BY-NC-ND 4.0 on the cited version[g]	baseline training and external test only; not redistributed

Table 15: Source datasets, distribution, and licenses as stated by each source. No single license satisfies all nine datasets simultaneously: three state no license, three impose non-commercial terms, and three are share-alike. AgroGround therefore partitions its release by source (see below). No source image or source text is redistributed, and the model trained on PlantSeg is not released.

Notes on the license column. [a] CDDM: the repository contains no LICENSE file and GitHub reports no license. The paper states an intent to release the dataset but names no terms. The CC BY-NC-ND 4.0 associated with the arXiv preprint is the preprint's submission license, a `cc-by-nc-4.0` tag on a third-party mirror was not set by the dataset authors, and the Tongyi Qianwen license inside the repository covers vendored Qwen-VL code. [b] LeafNet is tagged CC BY 4.0 and LeafBench MIT, while the LeafBench card's prose states CC BY 4.0 and the authors' LeafBench code repository carries a CC BY-NC 4.0 LICENSE file. The download gate for both requires agreeing to non-commercial use and restricts intended use to academic research or education. We treat both as non-commercial academic use. [c] AgroMind: no license file or metadata field in the authors' repositories and no license statement in the paper. The CC BY-SA 4.0 that circulates for AgroMind traces to static badge images in the repository READMEs and to a project-page footer that licenses the website. [d] AgroCoT: declared only as a Hugging Face dataset-card metadata field. The paper names no license and the repository has no LICENSE file. A third-party redistribution of the same data is tagged CC BY-NC 4.0. [e] AgroBench: no license in the repositories, the paper, or the supplementary material. Access is gated, so further terms may apply behind the gate. [f] AgMMU: CC BY-SA 4.0 is asserted in the paper from arXiv v2 onward and in the NeurIPS 2025 camera-ready. The released data carries only Hugging Face's unspecified `cc` tag and the code repository has no LICENSE file. [g] PlantSeg: the license differs by Zenodo version. We use the 15 September 2024 record (CC BY-NC-ND 4.0). Later versions variously carry CC BY-NC-ND 4.0, CC BY 4.0, or CC BY-NC 4.0.

Release. We release the AgroGround annotations (boxes, labels, prompts, and the evaluation set) keyed to source image identifiers, partitioned by source so that each shard carries terms compatible with its origin: CC BY-SA 4.0 for the MIRAGE-, AgroCoT-, and AgMMU-derived shards, CC BY-NC 4.0 for the LeafNet- and LeafBench-derived shards, and CC BY 4.0 for the remainder, whose sources state no license. The taxonomy, the construction and audit scripts, and the scoring code are released under Apache-2.0, with prediction files alongside the annotations. PlantSeg-derived artifacts are not redistributed. PlantSeg results are reproducible by applying the released scoring code to the cited Zenodo record under its own license, and the model trained on PlantSeg is not released. Hosting is by DOI with versioned releases.

R EXTENDED LIMITATIONS

Pseudo-label noise. Training boxes come from the annotator configuration, whose region IoU against human ground truth is 0.586 at precision 0.584. That the students reach 0.578–0.622 at precision 0.700–0.760 indicates they do not simply reproduce the teacher's boxes, but noise bounds fine-grained supervision quality, and the RL reward inherits it visibly, in the over-optimization degradation. Dense strata were identified via the automatic annotator's detection count and its pre-annotations seeded the human boxes (36.8% of final boxes are accepted detector proposals, kept essentially unchanged, Appendix C), so a mild ground-truth–detector correlation exists there. The human boxes nonetheless contain far more instances than the detector finds: the annotator configuration's recall against them is 0.295, and taxonomy prompts alone reach 0.411. **Duplication in the sources.** The sources copy from one another and from shared pools. Unique-file counts overstate diversity, and any evaluation on them without pixel-level deduplication risks contamination, as our own first benchmark version did. **Evaluation scope.** Ground truth reflects a single trained annotator

under written conventions with a removal-and-flag protocol, not multi-annotator consensus. The set covers 1,480 images and cannot cover all crop–disease combinations, and dense multi-lesion field images are scarce (180). **Benchmark contamination.** Seven of our eight sources are public evaluation sets or contain evaluation splits (CDDM's test split, MIRAGE, AgroMind, LeafBench, AgroBench, AgMMU, AgroCoT). Models trained on AgroGround are contaminated with respect to them. We release per-source and per-split flags so that any subset can be excluded. **Abstention is measured on leaf-centric imagery.** 492 of the 505 healthy benchmark images come from LeafNet, CDDM, and LeafBench, where the 12% model abstains on 95.9% and the RL model on 89.6%. The remaining 13 come from AgroMind and AgroCoT, where they abstain on 8 and 2 respectively. No healthy benchmark image comes from MIRAGE, AgroBench, or AgMMU, so the headline abstention rates are not established for field photography. **Input resolution.** Supervised fine-tuning caps images at 50,176 pixels while evaluation applies no cap, so every fine-tuned model is evaluated well above its training resolution (Appendix D). This plausibly contributes to under-enumeration and is an obvious lever for future work. **One backbone family.** All fine-tuning uses Qwen3-VL (2B and 8B). Seed variance is reported for the main supervised model (three seeds, Appendix N) but not for every configuration, and no second model family, closed-source zero-shot suite, or detector fine-tuned on the pseudo-labels is included. **Known model failures.** Under-enumeration on dense lesions remains the principal open problem (recall 0.362 after RL). Open-set naming is absent in all models. Abstention learned from healthy negatives does not transfer to wrong-target queries on diseased images (Appendix F), and the fine-tuned 8B loses the base's recognition advantage on AgroBench.