AGROVG: A LARGE-SCALE MULTI-SOURCE BENCHMARK FOR AGRICULTURAL VISUAL GROUNDING

Anonymous authors
Paper under double-blind review

ABSTRACT

Visual grounding, the task of localizing objects described by natural-language expressions, is a foundational capability for agricultural AI systems, enabling applications such as selective weeding, disease monitoring, and targeted harvesting. Reliable evaluation of agricultural visual grounding remains challenging because agricultural targets are often small, repetitive, occluded, or irregularly shaped, and instructions may refer to one, many, or no objects in an image. Evaluating this capability therefore requires jointly testing localization accuracy, target-set completeness, and existence-aware abstention. To address these challenges, we introduce **AgroVG**, a multi-source benchmark that formulates agricultural grounding as generalized set prediction: given an image and a referring expression, a model must return all matching target instances or abstain when no target is present. AgroVG contains 10,071 annotation-grounded image-query pairs from ten source datasets across six target families: crop/weed, fruit, wheat head, pest, plant disease, and tree canopy. It supports bounding-box grounding (T1) across all six families and instance-mask grounding (T2) on sources with reliable instance-level pixel annotations, with queries covering single-target, multi-target, and target-absent regimes. AgroVG further provides task-specific protocols for box-set matching and query-level mask coverage. Zero-shot evaluation of 26 model configurations spanning closed-source MLLMs, open-source VLMs, and specialized grounding systems reveals persistent gaps: the best multi-target Set-F_1 reaches only 0.35, and the best positive-query mask success rate at IoU@0.75 remains below 0.17. Data and code are available at `https://anonymous.4open.science/r/AgroVG-5172/`.

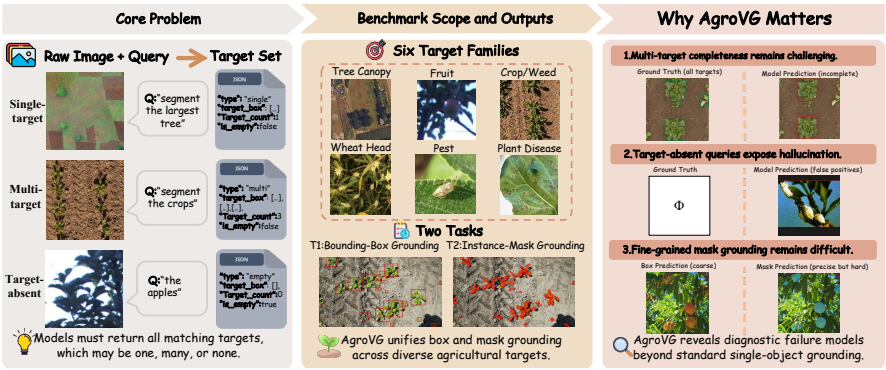


Figure 1: **AgroVG at a glance.** AgroVG casts agricultural grounding as generalized set prediction: a query over an agricultural image may require one target, multiple targets, or no output. The benchmark spans six target families and two output protocols, bounding boxes (T1) and instance masks (T2), to diagnose multi-target completeness, target-absent abstention, and mask grounding.

1 INTRODUCTION

Agricultural AI systems must ground human instructions in complex crop environments, from removing "all mature weeds among soybean plants" to localizing "every yellowing leaf in the upper canopy." Beyond image-level recognition, this requires identifying *which* objects satisfy a description, *where* they are, and *whether* they exist. Visual grounding, studied through referring expression

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.

comprehension and segmentation (Kazemzadeh et al., 2014; Mao et al., 2016), thus supports agricultural robots, precision monitoring, and vision-language-action pipelines (Vougioukas, 2019; Kim et al., 2024).

Despite rapid progress in visual grounding, existing benchmarks do not provide this evaluation setting for agricultural imagery. General-domain benchmarks have evolved from single-target referring expressions on everyday objects (Yu et al., 2016; Mao et al., 2016) to generalized settings with multi-target and target-absent queries (He et al., 2023; Liu et al., 2023), while remote-sensing benchmarks evaluate language-conditioned localization in top-down aerial imagery (Zhan et al., 2023; Yuan et al., 2024; Ding et al., 2024). These regimes differ substantially from close-range field, orchard, and canopy scenes. In agriculture, recent multimodal benchmarks evaluate recognition, visual question answering, captioning, and domain-specific reasoning (Zhou & Ryo, 2024; Shinoda et al., 2025; Gauba et al., 2025; Li et al., 2025b), but they generally do not test whether a model can localize exact target regions, retrieve all matching instances, or abstain when no such target exists. The closest related agricultural grounding benchmark, gRef-CW (Haghighat et al., 2026), supports box-level generalized grounding on CropAndWeed imagery (Steininger et al., 2023), but is restricted to a single source dataset, a single crop–weed sub-domain, and provides no instance-mask grounding evaluation. A unified benchmark that spans agricultural target families and jointly evaluates bounding-box and instance-mask grounding remains missing, as summarized in Table 1.

Agricultural grounding poses three challenges. *(i) Dense, repetitive instances:* frames may contain dozens of similar wheat heads, clustered fruits, or interleaved crop and weed plants (David et al., 2021; Hani et al., 2020; Steininger et al., 2023), so expressions need not identify a unique target. *(ii) Small, occluded, irregular targets:* pests, lesions, and leaves are often small or partially hidden, and many have irregular, non-convex boundaries (Wu et al., 2019; Wei et al., 2026), motivating instance masks for faithful pixel-level evaluation. *(iii) Target-absent expressions:* field instructions may name absent objects, requiring abstention. These properties motivate generalized set prediction (Liu et al., 2023): models return one, many, or zero matching instances, evaluated as boxes or masks.

We introduce **AgroVG**, comprising 10,071 annotation-grounded image–query pairs from 10 datasets across six target families: crop/weed, fruit, wheat head, pest, plant disease, and tree canopy. It supports T1 (bounding-box grounding) across all six families and T2 (instance-mask grounding) where reliable instance-level pixel annotations exist (Fig. 1). Both tasks cover single-target, multi-target, and target-absent queries, yielding six diagnostic cells. Every positive query maps to verified object identities and a target set; every target-absent query has a verified empty set under the benchmark annotations. We unify source annotations, apply expert quality control, and provide box-set matching and query-level mask coverage protocols (Sec. 4).

Zero-shot evaluation of 26 model configurations spanning closed-source MLLMs, open-source VLMs, and specialized grounding systems shows that AgroVG is far from saturated: even the best T1 multi-target configuration reaches only 0.35 Set-F_1, and the best T2 positive-query mask success rate remains below 0.17 at IoU@0.75. Models also show poor abstention calibration, either hallucinating boxes or masks on target-absent queries or over-abstaining on positive queries. These results position AgroVG as a challenging and diagnostic benchmark for agricultural vision-language grounding.

In summary, our contributions are:

- We formulate agricultural visual grounding as a generalized set-prediction problem and provide task-specific evaluation protocols for bounding-box grounding (T1) and instance-mask grounding (T2) across single-target, multi-target, and target-absent regimes.
- We construct AgroVG, a multi-source, multi-family benchmark with 10,071 annotation-grounded image–query pairs and verified object identities, covering six agricultural target families under quality-controlled annotations.
- We conduct a zero-shot evaluation of 26 representative configurations spanning closed-source MLLMs, open-source VLMs, and specialized grounding systems, revealing systematic failures in set completeness, existence-aware abstention, and fine-grained mask grounding.

Table 1: **Benchmark comparison.** AgroVG uniquely combines multi-source agricultural coverage with box and mask grounding across single-, multi-, and target-absent settings.

Benchmark	Year	Dom.	Size	#Src.	Target	BBox Grounding			Mask Grounding		
						Single	Multi	Absent	Single	Multi	Absent
RefCOCO (Yu et al., 2016)	2016	Nat.	19,994 img.	1	general	✓	✗	✗	✓	✗	✗
gRefCOCO (Liu et al., 2023)	2023	Nat.	19,994 img.	1	general	✗	✗	✗	✓	✓	✓
DIOR-RSVG (Zhan et al., 2023)	2023	RS	17,402 img.	1	aerial	✓	✗	✗	✗	✗	✗
RRSIS-D (Liu et al., 2024)	2024	RS	17,402 trip.	1	aerial	✗	✗	✗	✓	✗	✗
VRSBench (Ding et al., 2024)	2024	RS	29,614 img.	2	aerial	✓	✗	✗	✗	✗	✗
AgriBench (Zhou & Ryo, 2024)	2024	Agri.	1,784 img.	1	–	✗	✗	✗	✗	✗	✗
AgroBench (Shinoda et al., 2025)	2025	Agri.	4,218 img.	multi	–	✗	✗	✗	✗	✗	✗
AgroMind (Li et al., 2025b)	2025	Agri-RS	20,850 img.	10	–	✗	✗	✗	✗	✗	✗
AgMMU (Gauba et al., 2025)	2025	Agri.	–	–	–	✗	✗	✗	✗	✗	✗
gRef-CW (Haghighat et al., 2026)	2026	Agri.	8,034 img.	1	crop/weed	✓	✓	✓	✗	✗	✗
AgroVG (Ours)	**2026**	**Agri.**	**10,071 pairs**	**10**	**6 families**	✓	✓	✓	✓	✓	✓

Notes. Size follows each benchmark's reported unit: img. = images, trip. = triplets, and pairs = image–query pairs. Target denotes the domain of language-conditioned localization; "–" indicates no instance-level grounding protocol. AgroVG covers six agricultural target families: crop/weed, fruit, wheat head, pest, plant disease, and tree canopy.

2 Related Work

2.1 Visual Grounding Benchmarks

Visual grounding connects natural-language expressions to localized visual evidence and is commonly studied along two complementary axes: Referring Expression Comprehension (REC) with bounding boxes and Referring Expression Segmentation (RES) with masks (Hu et al., 2016; Yu et al., 2018; Deng et al., 2021). Natural-image benchmarks such as RefCOCO, RefCOCO+ (Yu et al., 2016), and RefCOCOg (Mao et al., 2016) established canonical single-target grounding protocols on everyday objects, where each expression is typically assumed to refer to exactly one target. More recent generalized grounding benchmarks, such as gRefCOCO and the GREC/GRES settings (He et al., 2023; Liu et al., 2023), relax this assumption by introducing multi-target and target-absent expressions, reformulating grounding as a set-prediction problem. Together, these efforts establish two evaluation axes, output granularity (box vs. mask) and target-set cardinality (single, multi, absent), that we adopt and instantiate for agricultural imagery in AgroVG.

Domain-specific grounding benchmarks have also emerged in remote sensing. *Box-level grounding* benchmarks include RSVG (Sun et al., 2022), DIOR-RSVG(Zhan et al., 2023), and subsequent datasets that extend referring expression comprehension to higher-resolution, optical, and drone-view imagery (Lan et al., 2024; Li et al., 2024; Liu et al., 2025a; Sun et al., 2025). *Mask-level referring segmentation* benchmarks extend language-conditioned localization to pixel-precise outputs (Yuan et al., 2024; Dong et al., 2025; Yang et al., 2025d; Ye et al., 2026). More recent work further scales this line toward large-scale, generalized, and reasoning-oriented remote-sensing grounding (Quenum et al., 2025; Li et al., 2025a; Xin et al., 2025; Jiang et al., 2025; Ni et al., 2026), as well as spatial and cross-view reasoning (Chu et al., 2025; Zhou et al., 2025). VRSBench (Ding et al., 2024) additionally combines object-level references with VQA and captioning. Despite this rapid progress, the remote-sensing regime, typically top-down imagery dominated by vehicles, ships, buildings, aircraft, and land-cover regions, differs fundamentally from close-range agricultural imagery in scale distributions, target densities, occlusion patterns, and viewpoint characteristics. This leaves agricultural visual grounding systematically under-evaluated.

2.2 Agricultural Vision-Language Benchmarks

Recent agricultural vision-language benchmarks have advanced the evaluation of multimodal models in agricultural settings (Yang et al., 2025c;b;a; Yan et al., 2026). AgriBench (Zhou & Ryo, 2024), AgroBench (Shinoda et al., 2025), AgroMind (Li et al., 2025b), and AgMMU (Gauba et al., 2025) evaluate agricultural vision-language models on tasks such as recognition, visual question answering, captioning, and domain-specific reasoning. While these benchmarks are important for measuring high-level agricultural understanding, they generally do not provide a unified instance-level protocol for language-conditioned localization. A model may correctly answer whether a disease

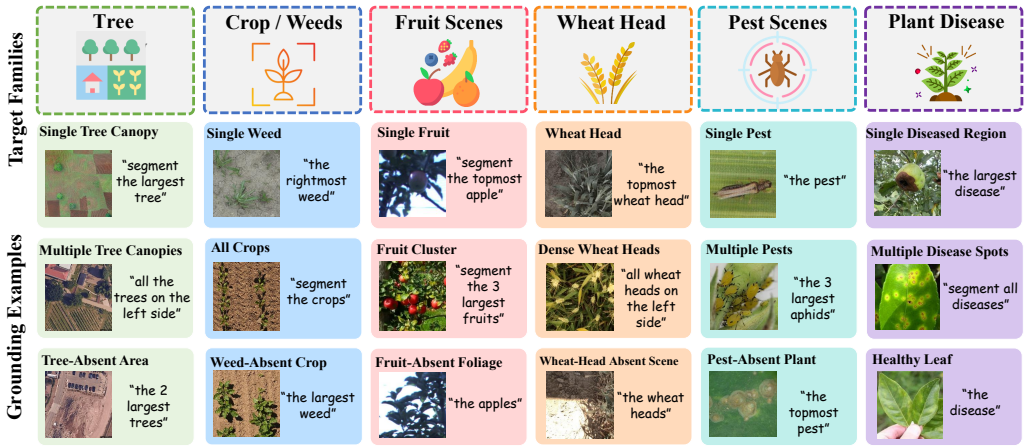


Figure 2: **Target families and grounding regimes in AgroVG.** AgroVG covers six target families (columns): tree canopy, crop/weed, fruit, wheat head, pest, and plant disease. For each family, we show queries under three regimes (rows): *single-target*, *multi-target*, and *target-absent*, where the output is one localized instance, a set of instances, or an empty prediction.

is present or describe a fruit scene, but this does not test whether it can localize the exact diseased regions, retrieve all valid target instances, or abstain when the referred target is absent.

Agricultural detection and segmentation datasets provide rich object-level supervision for crop–weed segmentation, fruit detection, plant phenotyping, pest recognition, disease segmentation, and tree canopy analysis. However, standard detection and segmentation are not equivalent to visual grounding: they predict instances from predefined categories, whereas grounding requires selecting the target set specified by a natural-language expression (Yu et al., 2016). The same image may require one instance, multiple instances, or no output depending on the query. The closest related agricultural grounding benchmark is gRef-CW (Haghighat et al., 2026), which provides box-level generalized referring expression comprehension for CropAndWeed imagery but no mask-grounding evaluation. As discussed in Sec.1 and summarized in Table 1, AgroVG complements this line of work along three dimensions: multi-source aggregation, multi-family coverage, and joint bounding-box and instance-mask evaluation.

3 AgroVG Benchmark

3.1 Task Formulation

AgroVG formulates agricultural visual grounding as a *generalized set-prediction problem*. Given an image I and a referring expression q, the goal is to localize all visual targets satisfying q. Let $\mathcal{O}(I)$ be the annotated instances in I. Each query induces a target set $\mathcal{G}(I,q) \subseteq \mathcal{O}(I)$ with $K = |\mathcal{G}(I,q)|$, where K is determined by both the image content and the query rather than fixed a priori. AgroVG therefore supports single-target, multi-target, and target-absent grounding under two complementary output protocols: bounding-box grounding (T1) and instance-mask grounding (T2).

T1: Bounding-box grounding. T1 represents each target instance by an axis-aligned box. The ground-truth output is $\mathcal{B}^*(I,q) = \{b(o) \mid o \in \mathcal{G}(I,q)\}$, where $b(o) = (x^{\min}, y^{\min}, x^{\max}, y^{\max})$, and a model predicts a box set $\hat{\mathcal{B}}(I,q)$. T1 evaluates box-level localization and aligns with the output format of grounding detectors and many grounding-capable MLLMs.

T2: Instance-mask grounding. T2 represents each target instance by a binary mask. The ground-truth output is $\mathcal{M}^*(I,q) = \{m(o) \mid o \in \mathcal{G}(I,q)\}$, where $m(o) \in \{0,1\}^{H \times W}$, and a model predicts a mask set $\hat{\mathcal{M}}(I,q)$. T2 evaluates pixel-precise grounding for agricultural targets with irregular shapes, partial occlusion, or dense overlap.

Query regimes. Following generalized grounding settings (Liu et al., 2023; He et al., 2023), AgroVG evaluates three regimes shared by T1 and T2: *single-target* queries with $K = 1$, *multi-target* queries with $K > 1$, and *target-absent* queries with $K = 0$, where no verified target instance

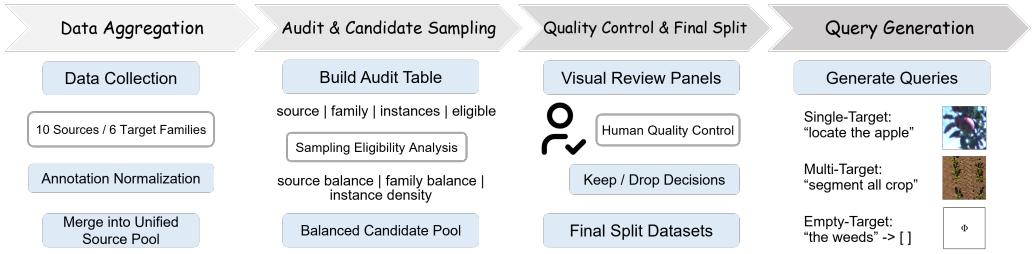


Figure 3: **AgroVG construction pipeline.** AgroVG converts heterogeneous agricultural datasets into verified image–query pairs through annotation normalization, audit-driven sampling, expert visual review, split assignment, and split-aware query generation. Positive and target-absent queries are generated after image verification, ensuring each query is tied to stable ground-truth annotations.

within the benchmark target space satisfies the expression. These regimes probe object localization, set completeness, and existence-aware abstention, yielding six $\langle \text{task}, \text{regime} \rangle$ evaluation cells. Fig.2 illustrates these regimes across the six target families with representative referring queries.

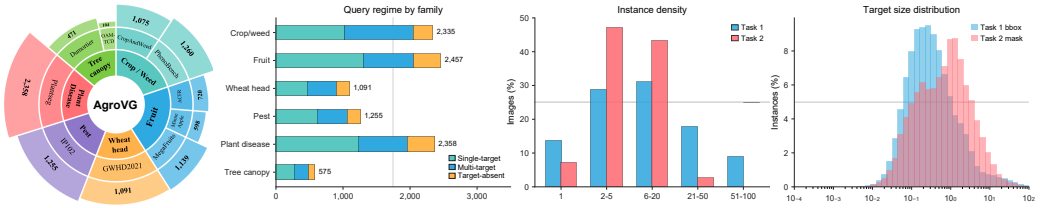
3.2 Benchmark Construction

As shown in Fig. 3, AgroVG follows a split-aware construction order: candidate images are first normalized, audited, reviewed, and assigned to final splits before referring query generation. This ensures each query is derived from a verified image and stable ground-truth annotations.

Data sources and annotation normalization. AgroVG is built from ten source datasets covering six target families: *crop/weed* (PhenoBench (Weyler et al., 2024), CropAndWeed (Steininger et al., 2023)), *fruit* (MinneApple (Hani et al., 2020), ACFR (Bargoti & Underwood, 2017), MegaFruits (Wang et al., 2026)), *wheat head* (GWHD (David et al., 2021)), *pest* (IP102 (Wu et al., 2019)), *plant disease* (PlantSeg (Wei et al., 2026)), and *tree canopy* (OAM-TCD (Veitch-Michaelis et al., 2024), Dumortier et al. (Dumortier, 2025)). T1 covers all six families using nine sources (all except MegaFruits). T2 covers crop/weed, fruit, and plant disease using five mask-capable sources: PhenoBench, CropAndWeed, PlantSeg, MegaFruits, and ACFR (Table A1). The unified instance-level schema records image metadata, target family, instance ID, bounding box, available mask references, and annotation provenance. We map source labels to six shared target families, excluding ambiguous mappings and retaining reliable fine-grained labels for compatible query templates. T2 masks require pixel-level evidence. Available native instance masks are used directly; otherwise, we derive masks from pixel-level annotations and instance cues, such as circle seeds, stem points, or polygons. AgroVG never synthesizes masks from bounding-box geometry alone. Detailed derivation rules and filtering thresholds are provided in Appendix B.2.

Audit-driven sampling and expert quality control. To balance sources of varying scale and density, an image-level *audit table* records source identity, target family, queryable instance counts, density buckets, and query-regime eligibility. Sampling balances source, family, and instance density, prioritizing multi-ready images for multi-target queries. Audit fields, source quotas, and density-bucket statistics are detailed in Appendix B.3. Review panels overlay normalized boxes or masks on the original RGB image. A team of ten domain experts reviewed T1 bounding boxes, T2 masks, and the associated category labels. A separate blind re-annotation audit of 53 derived-mask images yields instance-level mIoU of 0.888 against manual references (Appendix B.2). The retained images are partitioned into `dev` and `test` splits at the image level, stratified by source dataset and target family, with no image appearing in both splits.

Split-aware query generation. Queries are generated only from the approved final splits, ensuring that each query is linked to a verified image and a stable ground-truth target set. Each query record stores the image identity, query text, task type, query regime, and the associated target set: T1 records store target object IDs and target boxes, while T2 records store target instance IDs used to recover the query mask; for target-absent queries, the target set is empty. Queries are annotation-grounded and template-based, covering category-level and fine-grained references, geometric rank (largest, leftmost, etc.), spatial side selection, relative relations between target families, set-level quantification, and target-absent negatives. For T1, queries are target descriptions whose expected outputs are bounding-box sets; for T2, queries are segmentation-oriented expressions whose ex-



(a) (b) (c) (d)

Figure 4: **Dataset statistics of AgroVG.** Panels summarize source/family composition, query-regime counts, instance-density buckets, and target-size distributions for T1 and T2.

pected outputs are instance-mask sets. Generated queries are automatically verified for target-count consistency and ambiguity, and a stratified sample from each $\langle \text{source}, \text{regime} \rangle$ cell undergoes additional manual inspection. The full template library and verification rules are provided in Appendix B.4.

Dataset statistics. After the pipeline, AgroVG comprises 10,071 annotation-grounded image–query pairs over expert-reviewed agricultural images. T1 covers all six target families, while T2 is instantiated on the three families supported by reliable instance-level pixel annotations (crop/weed, fruit, and plant disease). As shown in Fig. 4, source contributions are bounded so that no single source or family dominates (a); each family preserves a meaningful target-absent share alongside positive queries (b); and both T1 and T2 cover a wide instance-density and target-size range, with most examples in moderately dense scenes of 2–20 instances per image (c,d), reflecting the multi-target nature of agricultural grounding. Compared with existing agricultural grounding resources, AgroVG uniquely combines multi-source and multi-family coverage with both bounding-box and instance-mask grounding under single-target, multi-target, and target-absent regimes.

4 Evaluation Protocol

We evaluate AgroVG under three principles consistent with its generalized grounding formulation: (i) *box-set matching* for T1, (ii) *query-level mask coverage* for T2, and (iii) *multi-granularity reporting* across query regimes, sources, target families, and program types. We define metrics separately for the two output protocols because T1 predicts instance-level bounding-box sets, whereas T2 predicts query-level binary masks induced by target instance IDs.

4.1 T1: Bounding-Box Set Matching

For each T1 query, the model prediction is parsed into a set of valid boxes $\hat{\mathcal{B}}(I, q)$ and compared with the ground-truth box set $\mathcal{B}^*(I, q)$. For each IoU threshold $\tau \in \{0.50, 0.75\}$, we construct a bipartite graph between ground-truth and predicted boxes, adding an edge whenever the pairwise box IoU is at least τ. We then compute a maximum-cardinality matching on this thresholded graph. The objective is to maximize the number of threshold-satisfying matches rather than the sum of matched IoUs.

Matched pairs are counted as TP_τ, unmatched predictions as FP_τ, and unmatched ground-truth boxes as FN_τ. We compute query-level precision, recall, and Set-F_1, and report Set-F_1 at $\tau \in \{0.50, 0.75\}$ as the primary T1 metric. We report macro aggregation as the primary T1 score, which averages query-level scores; micro aggregation is included in the released evaluator summaries.

For single-target queries, we additionally report a strict accuracy: a query is correct only when the model predicts exactly one valid box and that box matches the unique ground-truth box. For target-absent queries, the correct prediction is the empty set. Empty predictions are counted as correct for Empty Accuracy, while any predicted box on a target-absent query is counted as a false positive.

4.2 T2: Query-Level Mask Coverage

For T2, the ground-truth mask is generated from the image-level instance map and the query's target instance IDs. Pixels belonging to any target instance are foreground, yielding a query-level binary mask. The model output is also evaluated as a binary mask after thresholding. Although T2 is con-

Table 2: Main results on **AgroVG-T1**, grouped by model category. All/Fam denote overall/family-macro Set-F_1; S-Acc, M-F1, and E-Acc are single-, multi-, and empty-query diagnostics. **Bold** and underlined values indicate the best and second-best results.

Model	Set-F_1			Diagnostics		
	All@.5	Fam@.5	All@.75	S-Acc	M-F1	E-Acc
Closed-source MLLMs						
GPT-4o	0.1505	0.1512	0.1234	0.0451	0.0146	0.7990
GPT-5.4	0.2736	0.2612	0.1421	0.2287	0.1653	0.6683
Gemini 2.5 Pro	0.1849	0.1744	0.0793	0.1401	0.1583	0.3769
Claude Sonnet 4.6	0.1995	0.1900	0.0997	0.1691	0.1132	0.4975
Open-source VLMs						
DeepSeek-VL2-tiny	0.2227	0.2153	0.1817	0.1031	0.0616	**0.9749**
DeepSeek-VL2-small	0.3191	0.3057	0.2179	0.2995	0.2344	0.5528
InternVL3.5-8B	<u>0.3963</u>	<u>0.3816</u>	0.2387	<u>0.3430</u>	**0.3540**	0.5829
InternVL3.5-38B	**0.3978**	**0.3920**	**0.2569**	**0.3527**	<u>0.2980</u>	0.7588
Qwen3-VL-8B	0.3271	0.3174	0.2186	0.2415	0.1737	0.8894
Qwen3-VL-32B	0.3748	0.3673	<u>0.2393</u>	0.3188	0.2626	0.8090
Specialized grounding systems						
OWLv2-Large	0.2951	0.2819	0.2361	0.0805	0.1576	<u>0.9246</u>
Florence-2-Large	0.2019	0.1871	0.1326	0.2576	0.2034	0.0000
Grounding DINO 1.5	0.2471	0.2286	0.1659	0.1417	0.2910	0.0151

structed from instance-level annotations, evaluation measures query-level coverage of the referred target region rather than per-instance mask matching, avoiding the need to compare heterogeneous predictors by instance-level mask outputs.

For each positive query, we compute pixel-level intersection, union, ground-truth area, and predicted area, from which IoU and Dice are derived. We report query-averaged mIoU and mDice, as well as cumulative cIoU and cDice computed by summing intersections, unions, and foreground areas across positive queries before taking the ratio. mIoU treats each query equally, while cIoU reflects overall pixel-level coverage and is more influenced by large masks. We also report query-level success rates IoU@0.50 and IoU@0.75, using thresholded success rather than AP-style metrics because the evaluated models produce confidence scores with different calibrations or no confidence scores at all. For target-absent queries, the ground-truth mask is empty; an empty prediction is counted as correct for Empty Accuracy, while any predicted foreground is counted in the empty false-positive rate.

4.3 Reporting Granularity

Combining two output protocols and three query regimes yields six $\langle \text{task}, \text{regime} \rangle$ evaluation cells. Main results report the primary metrics for each cell: strict accuracy for T1 single-target queries, Set-F_1 at IoU thresholds for T1 multi-target queries, Empty Accuracy for target-absent queries, and mIoU/cIoU/IoU@τ for T2 positive queries. Each cell is broken down by source dataset, target family, query program type, and, for T2, mask provenance. Full precision, recall, Set-F_1, Dice, false-positive rates, and diagnostic breakdowns at the standard thresholds are provided in Appendix C.1.

5 Experiments

We evaluate existing multimodal and grounding systems zero-shot on the AgroVG `test` split under the protocols in Sec. 4: no model is trained or fine-tuned on AgroVG. The development split is used only for prompt selection, output-format design, and threshold tuning.

5.1 Evaluated Models and Setup

Evaluated models. We evaluate 26 task-specific configurations on AgroVG: 4 closed-source MLLMs, 6 open-source VLMs, and 16 specialized grounding configurations. The closed-source models are GPT-4o (OpenAI, 2024), GPT-5.4 (OpenAI, 2026), Gemini 2.5 Pro (Gemini Team et al.,

Table 3: Main results on **AgroVG-T2**, grouped by mask-grounding interface. All/Fam denote overall/family-macro mask scores; IoU@ thresholds are positive-query success rates, and E-Acc is empty-query accuracy. **Bold** and underlined values indicate the best and second-best results.

Model	Mask Quality			Success / Abstention		
	All-mIoU	Fam-mIoU	cIoU	IoU@.50	IoU@.75	E-Acc
Segmentation MLLMs / reasoning segmentation models						
LISA-7B	0.3088	0.3020	0.3869	**0.3085**	0.1542	0.0283
LISA-13B	**0.3177**	**0.3131**	0.3879	**0.3085**	0.1575	0.0755
GLaMM	0.2628	0.2540	0.2905	0.2206	0.1161	0.0377
PixelLM-7B	0.3080	0.2870	0.3829	0.2886	0.1111	0.0000
PixelLM-13B	0.2903	0.2706	0.3890	0.2687	0.1078	0.0189
GSVA (force-seg)	0.2159	0.2037	0.2252	0.1808	0.0896	0.0000
PSALM	0.2980	0.2759	**0.4151**	0.2935	**0.1675**	0.0000
Promptable / universal text-to-mask models						
X-Decoder	0.1470	0.1337	0.2313	0.1376	0.0796	0.3868
SEEM	0.1868	0.1739	0.2290	0.1857	0.1012	0.2925
Florence-2-Large	0.1875	0.1603	0.1866	0.1360	0.1078	0.1981
SAM 3	0.1161	0.0941	0.3112	0.0995	0.0680	**0.9340**
Two-stage grounding-to-mask references						
Grounded-SAM	0.2450	0.2430	0.2843	0.2322	0.1360	0.0094
OWLv2-Large+SAM	0.1042	0.1149	0.0825	0.0912	0.0498	0.6792

2023), and Claude Sonnet 4.6 (Anthropic, 2026). The open-source VLMs cover three families at two scales: DeepSeek-VL2 (tiny, small) (Wu et al., 2024), InternVL3.5 (8B, 38B) (Wang et al., 2025), and Qwen3-VL (8B, 32B) (Bai et al., 2025), all used for T1 box outputs. Specialized configurations include T1 grounding detectors, OWLv2-Large (Minderer et al., 2023), Florence-2-Large (Xiao et al., 2024), and Grounding DINO 1.5 (Liu et al., 2025b); T2 segmentation MLLMs, LISA (7B, 13B) (Lai et al., 2024), GLaMM (Rasheed et al., 2024), PixelLM (7B, 13B) (Ren et al., 2024b), GSVA (force-seg) (Xia et al., 2024), and PSALM (Zhang et al., 2025); and T2 promptable/two-stage text-to-mask methods, X-Decoder (Zou et al., 2023a), SEEM (Zou et al., 2023b), Florence-2-Large, SAM 3 (Carion et al., 2025), Grounded-SAM (Ren et al., 2024a), and OWLv2-Large+SAM. Two-stage methods are marked with +SAM in Table 3. Open-source and specialized models run on NVIDIA A100 GPUs, while closed-source models are queried via official APIs.

Experimental setup. For T1, each model receives an image and a referring expression and is instructed to return all matching target instances as a list of axis-aligned bounding boxes; for target-absent queries, the required output is an empty list. For T2, models with native mask outputs are queried directly using segmentation-oriented expressions, while two-stage methods pass predicted boxes to SAM as box prompts to obtain final masks. We use the same task-level prompt template whenever the model interface permits, with only minimal model-specific formatting adjustments. All outputs are processed by a unified pipeline that normalizes box coordinates to image-pixel space when needed, clips boxes to image bounds, discards malformed predictions, and scores predictions with the official AgroVG evaluation scripts. We use greedy or deterministic decoding whenever supported; otherwise, the provider default is recorded. Beyond these pre-specified parsing and validity rules, we do not apply test-set post-hoc filtering, re-ranking, ensembling, or manual correction. Full model lists, version identifiers, prompt templates, parsing rules, decoding settings, confidence thresholds, and hardware details are provided in Appendix C.2.

T1: Bounding-box grounding. Table 2 reports the main results on AgroVG-T1. Open-source VLMs achieve the strongest box-grounding performance among the evaluated model groups, with InternVL3.5-38B attaining the highest reported scores on the primary metrics and stricter localization, while InternVL3.5-8B scores highest on multi-target Set-F_1. This contrast suggests that stronger overall localization does not necessarily imply better set completion. Scaling improves both Qwen3-VL and DeepSeek-VL2, although the gains remain uneven across diagnostic axes.

Closed-source MLLMs and specialized grounding systems exhibit complementary failure modes rather than uniform dominance. GPT-5.4 is the strongest closed-source model but still trails InternVL3.5-38B, whereas GPT-4o shows stronger target-absent rejection but weaker positive-target localization. Among specialized systems, OWLv2-Large abstains well on target-absent queries,

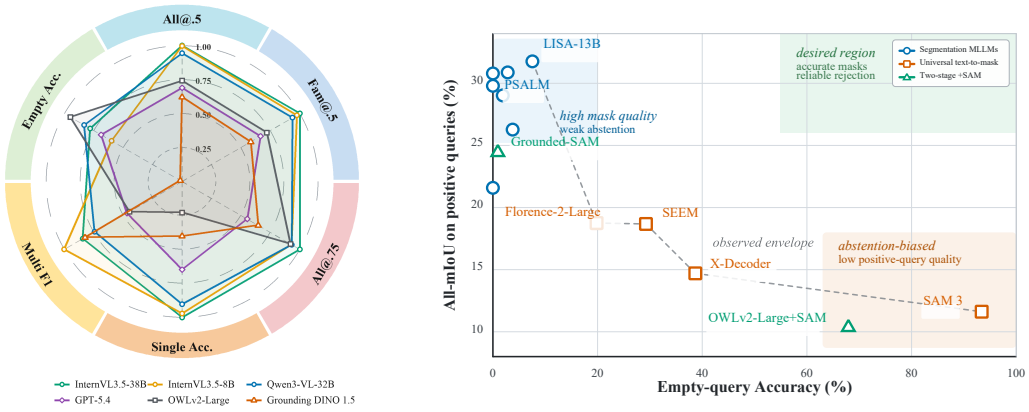


Figure 5: **Zero-shot diagnostics on AgroVG.** (a) T1 models across localization, set-completeness, and abstention; (b) T2 trade-off between positive-query mask quality and empty-query accuracy.

while Grounding DINO 1.5 and Florence-2-Large favor positive detections but abstain poorly. As shown in Fig.5(a), no single model dominates all T1 axes, confirming that AgroVG-T1 jointly tests localization accuracy, target-set completeness, and calibrated abstention.

T2: Instance-mask grounding. Table 3 summarizes the main results on AgroVG-T2. Segmentation MLLMs and reasoning segmentation models produce the strongest positive-query masks, with LISA-13B attaining the highest reported query-averaged and family-macro mIoU, and PSALM showing stronger cumulative coverage and stricter mask success. Nevertheless, success rates remain low, indicating that pixel-level agricultural grounding remains challenging for dense, irregular, and partially occluded targets.

The main T2 distinction is the trade-off between positive-query mask quality and target-absent rejection. SAM 3 achieves the strongest target-absent rejection but weak positive-query mask quality, whereas LISA-style and other mask-oriented models segment positive targets better but often hallucinate foreground when the target is absent. Two-stage systems show the same tension: Grounded-SAM improves positive-query mask coverage but rarely abstains, while OWLv2-Large+SAM abstains more reliably but produces weaker masks. As shown in Fig. 5(b), no model reaches the desired region where high positive-query mask quality coexists with reliable target-absent rejection.

Key findings. Across both tasks, AgroVG exposes three persistent gaps in current models. First, strong positive localization does not guarantee set completeness: the best T1 multi-target Set-F_1 is only 0.3540, indicating that models often miss valid target instances in dense agricultural scenes. Second, abstention is poorly calibrated: models with high Empty accuracy often behave conservatively, while models with stronger positive localization frequently hallucinate boxes or masks when the queried target is absent. Third, mask grounding remains the hardest protocol: the best T2 All-mIoU stays below 0.32 and the best IoU@.75 success rate remains below 0.17. Paired bootstrap intervals include zero for $4/12$ adjacent T1 comparisons and $7/12$ T2 comparisons, so close ranks warrant caution (Appendix D.4).

6 Conclusion

We introduce **AgroVG**, an agricultural visual grounding benchmark with 10,071 expert-verified pairs from ten source datasets spanning six target families. Annotation-grounded queries and task-specific protocols support bounding-box (T1) and instance-mask (T2) grounding under single-target, multi-target, and target-absent settings. Zero-shot evaluation of 26 model configurations reveals persistent gaps in multi-target completeness, existence-aware abstention, and pixel-level grounding of dense or irregular targets, motivating models with accurate localization and reliable abstention. AgroVG remains limited by template-based expressions and the availability of reliable instance-level pixel annotations, restricting T2 to three of the six target families. Future work will diversify queries, add mask-capable sources, and study controlled adaptation to agricultural imagery.

AI USE STATEMENT

We used generative AI tools to assist with translation, language editing, manuscript organization, LaTeX formatting, and critical feedback on methodological descriptions and the interpretation of experimental results. AI-generated writing suggestions were not treated as scientific evidence or cited as sources. The authors reviewed the AI-assisted revisions and take full responsibility for the final manuscript, including its scientific claims, reported results, and references.

ETHICS STATEMENT

AgroVG reuses ten public agricultural image datasets without collecting new imagery. Source provenance and attribution are retained, and source-specific access and redistribution terms remain applicable; where raw-image redistribution is restricted, users obtain images from the original sources (Appendices B.1 and E). Dataset construction includes expert review of bounding boxes, masks, and category labels (Appendix B.3). AgroVG is intended for research evaluation and does not certify readiness for autonomous agricultural intervention. Operational use requires additional validation under the intended crop, sensor, region, and field conditions. Potential misuse and effects on agricultural labor are discussed in Appendix A.

REPRODUCIBILITY STATEMENT

Sections 3–5 describe benchmark construction, task definitions, evaluation protocols, and experiments. Appendix B details the source datasets, annotation schema, mask derivation, query generation, quality control, and image-level development/test splits. Appendix C documents metric definitions, model identifiers, prompting and parsing rules, and the compute setup. Appendix D reports diagnostic results and uncertainty analyses, including paired bootstrap comparisons, human query rewriting, and repeated API evaluations. The anonymous repository is linked in the abstract; access and licensing are described in Appendix E. Repeating inference requires the corresponding source images, model checkpoints, or API access, and provider-side changes may affect regenerated API predictions.

REFERENCES

Anthropic. Introducing Claude Sonnet 4.6, 2026. URL https://www.anthropic.com/news/claude-sonnet-4-6.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, Mei Li, Kaixin Li, Zicheng Lin, Junyang Lin, Xuejing Liu, Jiawei Liu, Chenglong Liu, Yang Liu, Dayiheng Liu, Shixuan Liu, Dunjie Lu, Ruilin Luo, Chenxu Lv, Rui Men, Lingchen Meng, Xuancheng Ren, Xingzhang Ren, Sibo Song, Yuchong Sun, Jun Tang, Jianhong Tu, Jianqiang Wan, Peng Wang, Pengfei Wang, Qiuyue Wang, Yuxuan Wang, Tianbao Xie, Yiheng Xu, Haiyang Xu, Jin Xu, Zhibo Yang, Mingkun Yang, Jianxin Yang, An Yang, Bowen Yu, Fei Zhang, Hang Zhang, Xi Zhang, Bo Zheng, Humen Zhong, Jingren Zhou, Fan Zhou, Jing Zhou, Yuanzhi Zhu, and Ke Zhu. Qwen3-VL Technical Report, 2025. URL http://arxiv.org/abs/2511.21631.

Suchet Bargoti and James Underwood. Deep fruit detection in orchards. In *2017 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 3626–3633. IEEE, 2017. doi: 10.1109/ICRA.2017.7989417. URL http://ieeexplore.ieee.org/document/7989417/.

Nicolas Carion, Laura Gustafson, Yuan-Ting Hu, Shoubhik Debnath, Ronghang Hu, Didac Suris, Chaitanya Ryali, Kalyan Vasudev Alwala, Haitham Khedr, Andrew Huang, Jie Lei, Tengyu Ma, Arpit Kalla, Markus Marks, Joseph Greer, Meng Wang, Peize Sun, Roman Rädle, Effrosyni Mavroudi, Katherine Xu, Tsung-Han Wu, Yu Zhou, Liliane Momeni, Rishi Hazra, Shuangrui Ding, Sagar Vaze, Francois Porcher, Feng Li, Siyuan Li, Aishwarya Kamath, Ho Kei, Piotr Dollár, Nikhila Ravi, Kate Saenko, Pengchuan Zhang, and Christoph Feichtenhofer. SAM 3: Segment Anything with Concepts. 2025.

Meng Chu, Zhedong Zheng, Wei Ji, Tingyu Wang, and Tat-Seng Chua. Towards Natural Language-Guided Drones: GeoText-1652 Benchmark with Spatial Relation Matching. In Aleš Leonardis, Elisa Ricci, Stefan Roth, Olga Russakovsky, Torsten Sattler, and Gül Varol (eds.), *Computer Vision – ECCV 2024*, volume 15069, pp. 213–231. Springer Nature Switzerland, 2025. doi: 10.1007/978-3-031-73247-8_13. URL https://link.springer.com/10.1007/978-3-031-73247-8_13.

Etienne David, Mario Serouart, Daniel Smith, Simon Madec, Kaaviya Velumani, Shouyang Liu, Xu Wang, Francisco Pinto, Shahameh Shafiee, Izzat S.A. Tahir, Hisashi Tsujimoto, Shuhei Nasuda, Bangyou Zheng, Norbert Kirchgessner, Helge Aasen, Andreas Hund, Pouria Sadhegi-Tehran, Koichi Nagasawa, Goro Ishikawa, Sébastien Dandrifosse, Alexis Carlier, Benjamin Dumont, Benoit Mercatoris, Byron Evers, Ken Kuroki, Haozhou Wang, Masanori Ishii, Minhajul A. Badhon, Curtis Pozniak, David Shaner LeBauer, Morten Lillemo, Jesse Poland, Scott Chapman, Benoit De Solan, Frédéric Baret, Ian Stavness, and Wei Guo. Global Wheat Head Detection 2021: An Improved Dataset for Benchmarking Wheat Head Detection Methods. *Plant Phenomics*, 2021: 9846158, 2021. doi: 10.34133/2021/9846158. URL https://linkinghub.elsevier.com/retrieve/pii/S2643651524000591.

Jiajun Deng, Zhengyuan Yang, Tianlang Chen, Wengang Zhou, and Houqiang Li. TransVG: End-to-End Visual Grounding with Transformers. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 1749–1759. IEEE, 2021. doi: 10.1109/ICCV48922.2021.00179. URL https://ieeexplore.ieee.org/document/9710016/.

Jian Ding, Mohamed Elhoseiny, and Xiang Li. VRSBench: A Versatile Vision-Language Benchmark Dataset for Remote Sensing Image Understanding. In *Advances in Neural Information Processing Systems 37*, pp. 3229–3242. Neural Information Processing Systems Foundation, Inc. (NeurIPS), 2024. doi: 10.52202/079017-0106. URL http://www.proceedings.com/079017-0106.html.

Zhe Dong, Yuzhe Sun, Tianzhu Liu, Wangmeng Zuo, and Yanfeng Gu. Cross-Modal Bidirectional Interaction Model for Referring Remote Sensing Image Segmentation, 2025. URL http://arxiv.org/abs/2410.08613.

Jelle Dumortier. Annotated tree crown bounding boxes in urban/rural environment, 2025. URL https://zenodo.org/doi/10.5281/zenodo.15155081.

Aruna Gauba, Irene Pi, Yunze Man, Ziqi Pang, Vikram S Adve, and Yu-Xiong Wang. AGMMU: A Comprehensive Agricultural Multimodal Understanding Benchmark. 2025.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M. Dai, Anja Hauth, Katie Millican, et al. Gemini: A Family of Highly Capable Multimodal Models. *arXiv preprint arXiv:2312.11805*, 2023. doi: 10.48550/arXiv.2312.11805. URL https://arxiv.org/abs/2312.11805.

Mohammadreza Haghighat, Alzayat Saleh, and Mostafa Rahimi Azghadi. Multi-label Instance-level Generalised Visual Grounding in Agriculture, 2026. URL http://arxiv.org/abs/2603.06699.

Nicolai Hani, Pravakar Roy, and Volkan Isler. MinneApple: A Benchmark Dataset for Apple Detection and Segmentation. *IEEE Robotics and Automation Letters*, 5(2):852–858, 2020. doi: 10.1109/LRA.2020.2965061. URL https://ieeexplore.ieee.org/document/8954630/.

Shuting He, Henghui Ding, Chang Liu, and Xudong Jiang. GREC: Generalized Referring Expression Comprehension, 2023. URL http://arxiv.org/abs/2308.16182.

Ronghang Hu, Marcus Rohrbach, and Trevor Darrell. Segmentation from Natural Language Expressions. In Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling (eds.), *Computer Vision – ECCV 2016*, volume 9905, pp. 108–124. Springer International Publishing, 2016. doi: 10.1007/978-3-319-46448-0_7. URL http://link.springer.com/10.1007/978-3-319-46448-0_7.

Chengjie Jiang, Yunqi Zhou, Jiafeng Yan, Jing Li, Jiayang Li, Yue Zhou, Hongjie He, and Jonathan Li. GRASP: Geospatial pixel Reasoning viA Structured Policy learning, 2025. URL http://arxiv.org/abs/2508.17102.

Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara Berg. ReferItGame: Referring to Objects in Photographs of Natural Scenes. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 787–798. Association for Computational Linguistics, 2014. doi: 10.3115/v1/D14-1086. URL http://aclweb.org/anthology/D14-1086.

Moo Jin Kim, Karl Pertsch, Siddharth Karamcheti, Ted Xiao, Ashwin Balakrishna, Suraj Nair, Rafael Rafailov, Ethan Foster, Pannag Sanketi, Quan Vuong, Thomas Kollar, Benjamin Burchfiel, Russ Tedrake, Dorsa Sadigh, Sergey Levine, Percy Liang, and Chelsea Finn. OpenVLA: An Open-Source Vision-Language-Action Model. 2024.

Xin Lai, Zhuotao Tian, Yukang Chen, Yanwei Li, Yuhui Yuan, Shu Liu, and Jiaya Jia. LISA: Reasoning Segmentation via Large Language Model. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9579–9589. IEEE, 2024. doi: 10.1109/CVPR52733.2024.00915. URL https://ieeexplore.ieee.org/document/10658574/.

Meng Lan, Fu Rong, Hongzan Jiao, Zhi Gao, and Lefei Zhang. Language Query-Based Transformer With Multiscale Cross-Modal Alignment for Visual Grounding on Remote Sensing Images. *IEEE Transactions on Geoscience and Remote Sensing*, 62:1–13, 2024. doi: 10.1109/TGRS.2024.3407598. URL https://ieeexplore.ieee.org/document/10542207/.

Kaiyu Li, Zepeng Xin, Li Pang, Chao Pang, Yupeng Deng, Jing Yao, Guisong Xia, Deyu Meng, Zhi Wang, and Xiangyong Cao. SegEarth-R1: Geospatial Pixel Reasoning via Large Language Model, 2025a. URL http://arxiv.org/abs/2504.09644.

Ke Li, Di Wang, Haojie Xu, Haodi Zhong, and Cong Wang. Language-Guided Progressive Attention for Visual Grounding in Remote Sensing Images. *IEEE Transactions on Geoscience and Remote Sensing*, 62:1–13, 2024. doi: 10.1109/TGRS.2024.3423663. URL https://ieeexplore.ieee.org/document/10584552/.

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. 2025b.

Chang Liu, Henghui Ding, and Xudong Jiang. GRES: Generalized Referring Expression Segmentation. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 23592–23601. IEEE, 2023. doi: 10.1109/CVPR52729.2023.02259. URL https://ieeexplore.ieee.org/document/10204107/.

Junli Liu, Qizhi Chen, Zhigang Wang, Yiwen Tang, Yiting Zhang, Chi Yan, Dong Wang, Xuelong Li, and Bin Zhao. AerialVG: A Challenging Benchmark for Aerial Visual Grounding by Exploring Positional Relations. 2025a.

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 Aleš Leonardis, Elisa Ricci, Stefan Roth, Olga Russakovsky, Torsten Sattler, and Gül Varol (eds.), *Computer Vision – ECCV 2024*, volume 15105, pp. 38–55. Springer Nature Switzerland, 2025b. doi: 10.1007/978-3-031-72970-6_3. URL https://link.springer.com/10.1007/978-3-031-72970-6_3.

Sihan Liu, Yiwei Ma, Xiaoqing Zhang, Haowei Wang, Jiayi Ji, Xiaoshuai Sun, and Rongrong Ji. Rotated Multi-Scale Interaction Network for Referring Remote Sensing Image Segmentation. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 26648–26658. IEEE, 2024. doi: 10.1109/CVPR52733.2024.02517. URL https://ieeexplore.ieee.org/document/10657655/.

Junhua Mao, Jonathan Huang, Alexander Toshev, Oana Camburu, Alan Yuille, and Kevin Murphy. Generation and Comprehension of Unambiguous Object Descriptions. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11–20. IEEE, 2016. doi: 10.1109/CVPR.2016.9. URL http://ieeexplore.ieee.org/document/7780378/.

Matthias Minderer, Alexey Gritsenko, and Neil Houlsby. Scaling Open-Vocabulary Object Detection. 2023.

Shuo Ni, Di Wang, He Chen, Haonan Guo, Ning Zhang, and Jing Zhang. UniGeoSeg: Towards Unified Open-World Segmentation for Geospatial Scenes, 2026. URL http://arxiv.org/abs/2511.23332.

OpenAI. Hello GPT-4o, 2024. URL https://openai.com/index/hello-gpt-4o/.

OpenAI. GPT-5.4 Thinking System Card, 2026. URL https://openai.com/index/gpt-5-4-thinking-system-card/.

Jerome Quenum, Wen-Han Hsieh, Tsung-Han Wu, Ritwik Gupta, Trevor Darrell, and David M Chan. LISAT: Language-Instructed Segmentation Assistant for Satellite Imagery. 2025.

Hanoona Rasheed, Muhammad Maaz, Sahal Shaji, Abdelrahman Shaker, Salman Khan, Hisham Cholakkal, Rao M. Anwer, Eric Xing, Ming-Hsuan Yang, and Fahad S. Khan. GLaMM: Pixel Grounding Large Multimodal Model. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13009–13018. IEEE, 2024. doi: 10.1109/CVPR52733.2024.01236. URL https://ieeexplore.ieee.org/document/10655326/.

Tianhe Ren, Shilong Liu, Ailing Zeng, Jing Lin, Kunchang Li, He Cao, Jiayu Chen, Xinyu Huang, Yukang Chen, Feng Yan, Zhaoyang Zeng, Hao Zhang, Feng Li, Jie Yang, Hongyang Li, Qing Jiang, and Lei Zhang. Grounded SAM: Assembling Open-World Models for Diverse Visual Tasks, 2024a. URL https://arxiv.org/abs/2401.14159.

Zhongwei Ren, Zhicheng Huang, Yunchao Wei, Yao Zhao, Dongmei Fu, Jiashi Feng, and Xiaojie Jin. PixelLM: Pixel Reasoning with Large Multimodal Model. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 26364–26373. IEEE, 2024b. doi: 10.1109/CVPR52733.2024.02491. URL https://ieeexplore.ieee.org/document/10656606/.

Risa Shinoda, Nakamasa Inoue, Hirokatsu Kataoka, Masaki Onishi, and Yoshitaka Ushiku. AgroBench: Vision-Language Model Benchmark in Agriculture, 2025. URL http://arxiv.org/abs/2507.20519.

Daniel Steininger, Andreas Trondl, Gerardus Croonen, Julia Simon, and Verena Widhalm. The CropAndWeed Dataset: A Multi-Modal Learning Approach for Efficient Crop and Weed Manipulation. In *2023 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 3718–3727. IEEE, 2023. doi: 10.1109/WACV56688.2023.00372. URL https://ieeexplore.ieee.org/document/10030808/.

Yuxi Sun, Shanshan Feng, Xutao Li, Yunming Ye, Jian Kang, and Xu Huang. Visual Grounding in Remote Sensing Images. In *Proceedings of the 30th ACM International Conference on Multimedia*, pp. 404–412. ACM, 2022. doi: 10.1145/3503161.3548316. URL https://dl.acm.org/doi/10.1145/3503161.3548316.

Zhichao Sun, Yepeng Liu, Zhiling Su, Huachao Zhu, Yuliang Gu, Yuda Zou, Zelong Liu, Gui-Song Xia, Bo Du, and Yongchao Xu. RefDrone: A Challenging Benchmark for Referring Expression Comprehension in Drone Scenes, 2025. URL http://arxiv.org/abs/2502.00392.

Josh Veitch-Michaelis, Andrew Cottam, Daniella Schweizer, Eben N Broadbent, David Dao, Ce Zhang, Angelica Almeyda Zambrano, and Simeon Max. OAM-TCD: A globally diverse dataset of high-resolution tree cover maps. 2024.

Stavros G. Vougioukas. Agricultural Robotics. *Annual Review of Control, Robotics, and Autonomous Systems*, 2(1):365–392, 2019. doi: 10.1146/annurev-control-053018-023617. URL https://www.annualreviews.org/doi/10.1146/annurev-control-053018-023617.

Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, Zhaokai Wang, Zhe Chen, Hongjie Zhang, Ganlin Yang, Haomin Wang, Qi Wei, Jinhui Yin, Wenhao Li, Erfei Cui, Guanzhou Chen, Zichen Ding, Changyao Tian, Zhenyu Wu, Jingjing Xie, Zehao Li, Bowen Yang, Yuchen Duan, Xuehui Wang, Zhi Hou, Haoran Hao, Tianyi Zhang, Songze Li, Xiangyu Zhao, Haodong Duan, Nianchen Deng, Bin Fu, Yinan He, Yi Wang, Conghui He, Botian Shi, Junjun He, Yingtong Xiong, Han Lv, Lijun Wu, Wenqi Shao, Kaipeng Zhang, Huipeng Deng, Biqing Qi, Jiaye Ge, Qipeng Guo, Wenwei Zhang, Songyang Zhang, Maosong Cao, Junyao Lin, Kexian Tang, Jianfei Gao, Haian Huang, Yuzhe Gu, Chengqi Lyu, Huanze Tang, Rui Wang, Haijun Lv, Wanli Ouyang, Limin Wang, Min Dou, Xizhou Zhu, Tong Lu, Dahua Lin, Jifeng Dai, Weijie Su, Bowen Zhou, Kai Chen, Yu Qiao, Wenhai Wang, and Gen Luo. InternVL3.5: Advancing Open-Source Multimodal Models in Versatility, Reasoning, and Efficiency, 2025. URL http://arxiv.org/abs/2508.18265.

Yanan Wang, Zhenghao Fei, Ruichen Li, and Yibin Ying. Learn from foundation model: Fruit detection model without manual annotation. *Pattern Recognition*, 174:112799, 2026. doi: 10.1016/j.patcog.2025.112799. URL https://linkinghub.elsevier.com/retrieve/pii/S0031320325014621.

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. URL https://www.nature.com/articles/s41597-025-06513-4.

Jan Weyler, Federico Magistri, Elias Marks, Yue Linn Chong, Matteo Sodano, Gianmarco Roggiolani, Nived Chebrolu, Cyrill Stachniss, and Jens Behley. PhenoBench – A Large Dataset and Benchmarks for Semantic Image Interpretation in the Agricultural Domain. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(12):9583–9594, 2024. doi: 10.1109/TPAMI.2024.3419548. URL http://arxiv.org/abs/2306.04557.

Xiaoping Wu, Chi Zhan, Yu-Kun Lai, Ming-Ming Cheng, and Jufeng Yang. IP102: A Large-Scale Benchmark Dataset for Insect Pest Recognition. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8779–8788. IEEE, 2019. doi: 10.1109/CVPR.2019.00899. URL https://ieeexplore.ieee.org/document/8954351/.

Zhiyu Wu, Xiaokang Chen, Zizheng Pan, Xingchao Liu, Wen Liu, Damai Dai, Huazuo Gao, Yiyang Ma, Chengyue Wu, Bingxuan Wang, Zhenda Xie, Yu Wu, Kai Hu, Jiawei Wang, Yaofeng Sun, Yukun Li, Yishi Piao, Kang Guan, Aixin Liu, Xin Xie, Yuxiang You, Kai Dong, Xingkai Yu, Haowei Zhang, Liang Zhao, Yisong Wang, and Chong Ruan. DeepSeek-VL2: Mixture-of-Experts Vision-Language Models for Advanced Multimodal Understanding. 2024.

Zhuofan Xia, Dongchen Han, Yizeng Han, Xuran Pan, Shiji Song, and Gao Huang. GSVA: Generalized Segmentation via Multimodal Large Language Models. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3858–3869. IEEE, 2024. doi: 10.1109/CVPR52733.2024.00370. URL https://ieeexplore.ieee.org/document/10658546/.

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 *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4818–4829. IEEE, 2024. doi: 10.1109/CVPR52733.2024.00461. URL https://ieeexplore.ieee.org/document/10657519/.

Zepeng Xin, Kaiyu Li, Luodi Chen, Wanchen Li, Yuchen Xiao, Hui Qiao, Weizhan Zhang, Deyu Meng, and Xiangyong Cao. SegEarth-R2: Towards Comprehensive Language-guided Segmentation for Remote Sensing Images, 2025. URL http://arxiv.org/abs/2512.20013.

Lian Yan, Haotian Wang, Chen Tang, Haifeng Liu, Tianyang Sun, Liangliang Liu, Yi Guan, and Jingchi Jiang. AgriEval: A Comprehensive Chinese Agricultural Benchmark for Large Language Models. *Proceedings of the AAAI Conference on Artificial Intelligence*, 40(40):34205–34213, 2026. doi: 10.1609/aaai.v40i40.40716. URL https://ojs.aaai.org/index.php/AAAI/article/view/40716.

Bo Yang, Yunkui Chen, Lanfei Feng, Yu Zhang, Xiao Xu, Jianyu Zhang, Nueraili Aierken, Runhe Huang, Hongjian Lin, Yibin Ying, and Shijian Li. AgriGPT-VL: Agricultural Vision-Language Understanding Suite, 2025a. URL http://arxiv.org/abs/2510.04002.

Bo Yang, Lanfei Feng, Yunkui Chen, Yu Zhang, Jianyu Zhang, Xiao Xu, Nueraili Aierken, and Shijian Li. AgriGPT-Omni: A Unified Speech-Vision-Text Framework for Multilingual Agricultural Intelligence, 2025b. URL http://arxiv.org/abs/2512.10624.

Bo Yang, Yu Zhang, Lanfei Feng, Yunkui Chen, Jianyu Zhang, Xiao Xu, Nueraili Aierken, Yurui Li, Yuxuan Chen, Guijun Yang, Yong He, Runhe Huang, and Shijian Li. AgriGPT: A Large Language Model Ecosystem for Agriculture, 2025c. URL http://arxiv.org/abs/2508.08632.

Zhigang Yang, Huiguang Yao, Linmao Tian, Xuezhi Zhao, Qiang Li, and Qi Wang. A Large-Scale Referring Remote Sensing Image Segmentation Dataset and Benchmark, 2025d. URL http://arxiv.org/abs/2506.03583.

Kai Ye, YingShi Luan, Zhudi Chen, Guangyue Meng, Pingyang Dai, and Liujuan Cao. RIS-LAD: A Benchmark and Model for Referring Image Segmentation in Low-Altitude Drone Imagery. 2026.

Licheng Yu, Patrick Poirson, Shan Yang, Alexander C. Berg, and Tamara L. Berg. Modeling Context in Referring Expressions. In Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling (eds.), *Computer Vision – ECCV 2016*, volume 9906, pp. 69–85. Springer International Publishing, 2016. doi: 10.1007/978-3-319-46475-6_5. URL http://link.springer.com/10.1007/978-3-319-46475-6_5.

Licheng Yu, Zhe Lin, Xiaohui Shen, Jimei Yang, Xin Lu, Mohit Bansal, and Tamara L. Berg. MAttNet: Modular Attention Network for Referring Expression Comprehension. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1307–1315. IEEE, 2018. doi: 10.1109/CVPR.2018.00142. URL https://ieeexplore.ieee.org/document/8578240/.

Zhenghang Yuan, Lichao Mou, Yuansheng Hua, and Xiao Xiang Zhu. RRSIS: Referring Remote Sensing Image Segmentation. *IEEE Transactions on Geoscience and Remote Sensing*, 62:1–12, 2024. doi: 10.1109/TGRS.2024.3369720. URL https://ieeexplore.ieee.org/document/10458079/.

Yang Zhan, Zhitong Xiong, and Yuan Yuan. RSVG: Exploring Data and Models for Visual Grounding on Remote Sensing Data. *IEEE Transactions on Geoscience and Remote Sensing*, 61:1–13, 2023. doi: 10.1109/TGRS.2023.3250471. URL https://ieeexplore.ieee.org/document/10056343/.

Zheng Zhang, Yeyao Ma, Enming Zhang, and Xiang Bai. PSALM: Pixelwise SegmentAtion with Large Multi-modal Model. In Aleš Leonardis, Elisa Ricci, Stefan Roth, Olga Russakovsky, Torsten Sattler, and Gül Varol (eds.), *Computer Vision – ECCV 2024*, volume 15092, pp. 74–91. Springer Nature Switzerland, 2025. doi: 10.1007/978-3-031-72754-2_5. URL https://link.springer.com/10.1007/978-3-031-72754-2_5.

Yue Zhou, Mengcheng Lan, Xiang Li, Litong Feng, Yiping Ke, Xue Jiang, Qingyun Li, Xue Yang, and Wayne Zhang. GeoGround: A Unified Large Vision-Language Model for Remote Sensing Visual Grounding, 2025. URL http://arxiv.org/abs/2411.11904.

Yutong Zhou and Masahiro Ryo. AgriBench: A Hierarchical Agriculture Benchmark for Multimodal Large Language Models, 2024. URL http://arxiv.org/abs/2412.00465.

Xueyan Zou, Zi-Yi Dou, Jianwei Yang, Zhe Gan, Linjie Li, Chunyuan Li, Xiyang Dai, Harkirat Behl, Jianfeng Wang, Lu Yuan, Nanyun Peng, Lijuan Wang, Yong Jae Lee, and Jianfeng Gao. Generalized Decoding for Pixel, Image, and Language. 2023a.

Xueyan Zou, Jianwei Yang, Hao Zhang, Feng Li, Linjie Li, Jianfeng Wang, Lijuan Wang, Jianfeng Gao, and Yong Jae Lee. Segment Everything Everywhere All at Once. 2023b.

AgroVG: A Large-Scale Multi-Source Benchmark for Agricultural Visual Grounding

(Supplementary Material)

Appendix Contents

A Broader Impacts 17

B Dataset Construction Details 17

B.1 Source Datasets, Task Coverage, and Licenses 17

B.2 Annotation Schema, Mask Derivation, and Validation 18

B.3 Sampling, Quality Control, and Splits 20

B.4 Query Generation and Dataset Statistics 21

C Evaluation Setup and Reproducibility 25

C.1 Metric Definitions and Edge Cases 25

C.2 Evaluated Models, Prompts, and Parsing 27

C.3 Compute and Reproducibility Artifacts 28

D Additional Results and Diagnostics 30

D.1 T1 Diagnostic Results 30

D.2 T2 Diagnostic Results 31

D.3 Qualitative Failure Patterns and Case Studies 33

D.4 Uncertainty in Model Comparisons 36

E Hosting, Licensing, and Maintenance 38

A Broader Impacts

AgroVG is intended to support research on reliable language-conditioned localization in plant-centric agricultural imagery. By aggregating ten source datasets across crop/weed, fruit, wheat head, pest, plant disease, and tree canopy targets, and by evaluating bounding-box and instance-mask grounding under single-target, multi-target, and target-absent regimes, AgroVG provides a standardized testbed for capabilities that are important for precision agriculture, instruction-following agricultural robots, selective harvesting, targeted weed control, and early disease or pest monitoring. Its diagnostic metrics highlight not only localization accuracy, but also multi-target completeness, existence-aware abstention, and pixel-level mask quality.

The benchmark also surfaces failure modes that are important to identify before deployment. Our zero-shot evaluation shows that current models remain weak in multi-target set completion, abstention calibration, and fine-grained mask grounding. In real agricultural systems, such failures could lead to incorrect interventions, such as treating healthy plants as diseased, missing parts of a weed cluster, or acting on a target that is absent from the observed scene. AgroVG should therefore be viewed as an evaluation benchmark rather than evidence of deployment readiness. Models evaluated on AgroVG require additional validation under the target crop variety, sensor configuration, geographic region, and field condition before use in operational decision-making.

Potential misuse should also be considered. Although AgroVG contains no human subjects, no facial imagery, and no personally identifiable information, improvements in generalized grounding could transfer to monitoring applications outside agriculture, including human or property surveillance. The AgroVG release is therefore intended for agricultural research use, and should follow the licenses, access restrictions, and intended-use conditions of the contributing source datasets. As with other agricultural automation technologies, downstream applications may also affect labor demand in scouting, weeding, harvesting, and crop monitoring; such impacts should be considered alongside technical performance, especially in regions where agricultural labor markets are sensitive to mechanization.

To mitigate these risks, we encourage users to report disaggregated results by source dataset, target family, and query regime; to validate models with additional regional data before deployment; and to avoid treating AgroVG scores as a certification of readiness for autonomous agricultural intervention. The benchmark is best used as a scientific tool for developing more reliable, instruction-aware, and existence-aware agricultural grounding systems.

B Dataset Construction Details

B.1 Source Datasets, Task Coverage, and Licenses

AgroVG aggregates ten public agricultural datasets covering six target families: crop/weed, fruit, wheat head, pest, plant disease, and tree canopy. We did not collect new imagery. Instead, we convert mature agricultural detection and segmentation resources into a unified grounding benchmark with annotation-grounded referring queries. This design improves reproducibility and allows AgroVG to cover a broader set of agricultural subdomains than a newly collected single-source dataset.

Sources are selected when they satisfy three conditions: they depict plant-centric agricultural targets, provide instance-level boxes or pixel-level evidence from which instance-level supervision can be obtained, and are publicly accessible under terms compatible with academic benchmarking. T1 uses sources with reliable instance-level localizations and covers all six target families. T2 imposes the stricter requirement of consistent instance-level pixel evidence under the AgroVG normalization protocol; MinneApple and OAM-TCD are used only for T1 in the current release because their segmentation annotations were not included in the audited T2 pool, which is a coverage limitation rather than a fundamental restriction. Table A1 summarizes source coverage, annotation evidence, task usage, and access information.

AgroVG does not override source-dataset licenses. Each record retains source provenance and attribution. When source terms restrict raw-image redistribution, AgroVG releases metadata, preprocessing scripts, and query annotations, and users obtain raw images from the original source. Full per-source license statements are provided in the released dataset card.

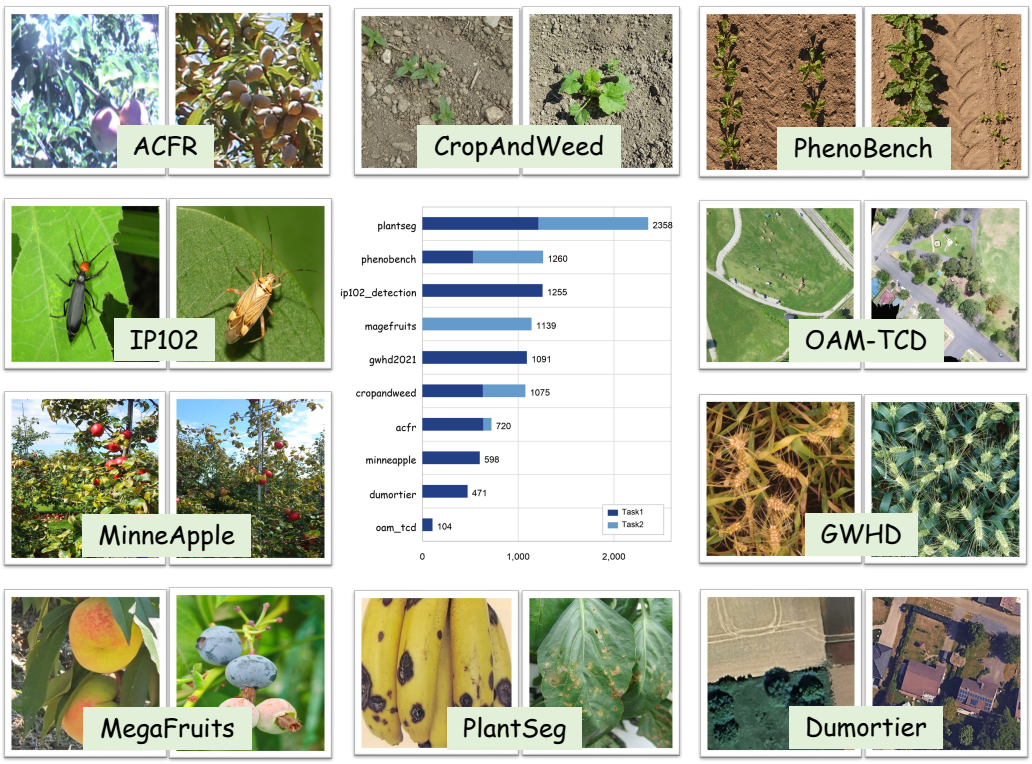


Figure A1: **Source datasets in AgroVG.** Representative examples from the ten agricultural source datasets are shown around the center plot. The center bar chart reports the number of image–query pairs contributed by each source, split by T1 bounding-box grounding and T2 instance-mask grounding.

Table A1: **Source datasets in AgroVG.** For each source we report the target family, native annotation evidence, AgroVG task usage, mask provenance for T2, native resolution, and access location. *Derived* masks are produced by the AgroVG normalization pipeline from source-side pixel evidence and instance cues.

Source	Year	Family	Native annotation	T1	T2	Mask prov.	Resolution	Access
PhenoBench (Weyler et al., 2024)	2024	crop/weed	instance masks	✓	✓	native	1024×1024	`phenobench.org`
CropAndWeed (Steininger et al., 2023)	2023	crop/weed	boxes + semantic + stem pts.	✓	✓	derived	high-res	`github.com/cropandweed`
MinneApple (Hani et al., 2020)	2020	fruit	polygonal masks	✓	—	—	high-res	`rsn.umn.edu/MinneApple`
ACFR (Bargoti & Underwood, 2017)	2016	fruit	foreground + circle seeds	✓	✓	derived	308×202	`data.acfr.usyd.edu.au`
MegaFruits (Wang et al., 2026)	2026	fruit	instance masks	—	✓	native	variable	`kaggle.com/mmwang0`
GWHD 2021 (David et al., 2021)	2021	wheat head	bounding boxes	✓	—	—	1024×1024	`global-wheat.com`
IP102 (Wu et al., 2019)	2019	pest	bounding boxes	✓	—	—	variable	`github.com/xpwu95/IP102`
PlantSeg (Wei et al., 2026)	2026	plant disease	polygons + binary mask	✓	✓	derived	variable	`zenodo.org/13762907`
OAM-TCD (Veitch-Michaelis et al., 2024)	2024	tree canopy	instance + semantic masks	✓	—	—	2048×2048	`zenodo.org/11617167`
Dumortier et al. (Dumortier, 2025)	2025	tree canopy	instance annotations	✓	—	—	640×640	`zenodo.org/15155081`

B.2 Annotation Schema, Mask Derivation, and Validation

The ten source datasets use different category vocabularies, coordinate conventions, annotation formats, and metadata structures. AgroVG normalizes them into a unified instance-level representation with three record layers (image, instance, query) summarized in Table A2; the full JSONL schema is released with the dataset metadata. T1 target sets are represented by object IDs and boxes; T2 target sets additionally store instance-map IDs; target-absent queries store an empty target set.

Bounding boxes use absolute-pixel `xyxy` coordinates with the origin at the upper-left image corner; T2 masks use uint16 instance maps aligned with the native image grid. After normalization, all records are subjected to automated validity checks, including valid dimensions, non-degenerate boxes, in-bound coordinates, non-empty masks, and box–mask consistency where applicable. Records failing these checks are removed or excluded from query generation.

Table A2: **Conceptual AgroVG annotation schema.** The released metadata contains the full field-level schema; this table summarizes the three record layers used for grounding evaluation.

Layer	Essential information	Purpose
Image	Source, split, image size, group ID, RGB path, and T2 instance-map path when available.	Source traceability and leakage-aware split control.
Instance	Object ID, target family, source class, bounding box, mask ID when available, and annotation provenance.	Defines localizable agricultural targets with auditable origin.
Query	Query text, task, regime, program type, target IDs, target boxes, and T2 target mask IDs.	Defines the annotation-grounded answer set for grounding evaluation.

Mask provenance. AgroVG-T2 uses native instance masks when available and derives masks only when the source provides pixel-level evidence. PhenoBench and MegaFruits provide native instance masks and are normalized directly. ACFR, CropAndWeed, and PlantSeg require source-specific derivation from pixel-level evidence and instance cues. AgroVG never synthesizes masks by filling bounding boxes, since box-filled masks would overestimate irregular agricultural targets and confound pixel-level evaluation. Each derived instance stores its mask provenance in the released metadata. All native and derived masks pass source-specific minimum-area and minimum-side filters (Table A3) before entering the T2 pool, with thresholds reflecting source-specific target scales.

ACFR-derived masks. ACFR provides binary fruit foreground masks together with circle annotations for individual fruit instances. We treat the binary foreground as the mask domain and use circle centers as instance seeds. Marker-based watershed on the distance transform partitions the foreground into instance regions. When a seed falls outside the foreground, it is projected to the nearest foreground pixel; when foreground or seed annotations are incomplete, we use conservative fallback rules based on annotation disks or connected components. Images without usable pixel-level fruit evidence are excluded from T2 rather than approximated from boxes.

CropAndWeed-derived masks. CropAndWeed provides semantic crop/weed masks, bounding boxes, and stem-point annotations. We process each semantic class independently: the class-specific foreground defines the mask domain, while stem points serve as instance seeds, with box-center fallback when a stem point is outside the foreground. Marker-based watershed partitions the semantic region into candidate instances. To reduce leakage between neighboring plants, each candidate is trimmed by its corresponding box region and reduced to the connected component containing the seed. This produces instance masks grounded in semantic pixels and instance cues rather than box geometry alone.

PlantSeg-derived masks. PlantSeg provides polygon annotations for disease regions, together with binary disease-foreground masks. We rasterize valid polygons into instance maps and preserve source split consistency during normalization. When polygon annotations are unavailable but a non-empty binary disease mask exists, we fall back to connected components on the binary foreground. If neither usable polygons nor binary foreground are available, the image is dropped from T2. Thus PlantSeg masks are derived from polygonal or pixel-level disease evidence, never from bounding boxes.

Derived-mask validation. We assess derived-mask fidelity on a pre-specified sample of 53 positive-query T2 test images from ACFR, CropAndWeed, and PlantSeg, stratified by instance density and enriched for touching instances. One annotator delineates all instances from the raw RGB images without viewing the derived masks. The sample, evaluation code, and LISA-13B reference model are fixed beforehand; existing model predictions remain frozen. This pixel-level audit is separate from the expert review described in Appendix B.3.

Table A4 reports agreement with the manual references. Instance-level mIoU is 0.888. At the query level, the manual masks are scored against the evaluated benchmark masks using the unchanged T2 protocol, yielding agreement of 0.886, compared with 0.257 for LISA-13B on the same subset. These results provide evidence of substantial model error alongside source-dependent annotation uncertainty, with lower agreement on ACFR. The single-annotator reference and enriched

Table A3: **T2 mask provenance, derivation rules, and filtering thresholds.** $A_{\min}$ is the minimum mask area in pixels; $S_{\min}$ is the minimum side length of the tight mask box. Derived masks are produced from pixel-level evidence and instance cues, never by filling bounding boxes.

Source	Provenance	Normalization or derivation rule	$A_{\min}$	$S_{\min}$
PhenoBench (Weyler et al., 2024)	Native	Native plant instance masks are normalized and filtered.	64	36
MegaFruits (Wang et al., 2026)	Native	Native fruit instance masks are normalized and filtered.	16	4
ACFR (Bargoti & Underwood, 2017)	Derived	Binary fruit foreground is partitioned using circle-center seeds.	8	2
CropAndWeed (Steininger et al., 2023)	Derived	Semantic crop/weed foreground is partitioned using stem-point or box-center cues, followed by cleanup.	24	36
PlantSeg (Wei et al., 2026)	Derived	Disease-region polygons are rasterized into instance masks; connected-component fallback is used only when polygons are unavailable but binary foreground exists.	4	2

Instances failing either threshold are excluded from the T2 pool. Rejection counts are reported in the released per-source statistics.

sample quantify agreement on this subset; they do not establish a calibrated performance ceiling or a benchmark-wide noise bound.

Table A4: **Derived-mask fidelity on 53 T2 test images.** One blind annotator redraws the instances. Query agreement scores the manual reference against evaluated benchmark masks; LISA-13B uses frozen predictions on the same positive queries.

Source	Images	Instance mIoU	Query agreement	LISA-13B query mIoU
ACFR	15	0.727	0.730	0.143
CropAndWeed	24	0.937	0.988	0.356
PlantSeg	14	0.909	0.878	0.208
Overall	53	0.888	0.886	0.257

B.3 Sampling, Quality Control, and Splits

The source datasets differ substantially in scale, target density, annotation provenance, and original task formulation. Uniform sampling from the merged source pool would over-represent the largest contributors and under-represent low-density, multi-target, and target-absent regimes. AgroVG therefore applies audit-driven sampling and visual review before query generation.

Audit table and candidate sampling. Source-side normalization outputs are merged into a unified pool with globally unique image and instance identifiers, while preserving source-side group identifiers (e.g., repeated captures of the same plant) for leakage-aware split construction. For each image, AgroVG computes an audit row recording source identity, target family, queryable instance count, instance-density bucket, mask provenance when applicable, and eligibility for single-target, multi-target, and target-absent query construction. A balanced T2 review pool of $N_{\text{review}} = 5{,}000$ images is drawn from this audit table under three controls: source-aware quotas (preventing source dominance and approximately balancing the three T2 families), multi-ready preference (ensuring sufficient multi-target query coverage), and density stratification (preserving sparse, moderate, and dense scenes). The resulting quotas and retained counts are reported in Table A5.

Expert visual review. Each sampled image is rendered as an RGB-overlay panel showing normalized boxes for T1 or instance masks and boundaries for T2. Following initial screening, a team of ten experts conducted an additional review of both tasks, checking bounding-box localization, mask

quality, and category labels. Review criteria include incomplete annotations, geometric misalignment, merged or over-split instances, and incorrect category labels.

Split construction. Review-approved images form the final image pools. For T2, 3,545 of the 5,000 reviewed candidates are retained. The retained images are partitioned into `dev` and `test` splits with an approximately $80\%/20\%$ ratio, applied per source. When source metadata defines groups, all images sharing a `group_id` are assigned to the same split to avoid leakage. Query generation is performed only after split assignment and is described in Appendix B.4.

Table A5: **T2 sampling and quality-control summary.** Review quota is the number of sampled candidate images before visual review. Retained images form the final T2 pool. Pass rates reflect source-specific differences in target density, annotation provenance, and mask reliability.

Source	Family	Review quota	Retained	Pass rate	Preferred density buckets
PhenoBench	crop/weed	900	732	81.3%	{1, 2–5, 6–20, 21–50}
CropAndWeed	crop/weed	800	443	55.4%	{1, 2–5, 6–20, 21–50}
MegaFruits-blueberry	fruit	700	554	79.1%	{2–5, 6–20, 21–50}
MegaFruits-peach	fruit	700	585	83.6%	{2–5, 6–20, 21–50}
ACFR-apples	fruit	300	84	28.0%	{1, 2–5, 6–20, 21–50}
PlantSeg	plant disease	1,600	1,147	71.7%	{1, 2–5, 6–20, 21–50}
Total	—	**5,000**	**3,545**	**70.9%**	—

B.4 Query Generation and Dataset Statistics

Queries are generated only after image normalization, visual review, and split assignment. This split-aware order ensures that each released query is tied to stable ground-truth annotations and that no query crosses dev/test image boundaries. AgroVG comprises 10,071 released image–query pairs (6,526 for T1 and 3,545 for T2), with query-type budgets of approximately 50% single-target, 35% multi-target, and 15% target-absent. T2 follows a one-query-per-image policy on its 3,545 final images, yielding a 1:1 correspondence; T1 releases 6,526 selected eligible image–query pairs from a larger expert-verified image pool of 6,664 images.

Design principles. The template library is governed by three principles. *(i) Annotation-grounded:* every emitted query carries an explicit target set traceable to verified instance annotations; positive queries store target object IDs and boxes (T1) or instance IDs (T2), and target-absent queries are verified to have an empty target set under the global family space. *(ii) Deterministic and program-typed:* each query is produced by exactly one named template (its `program_type`, stored in the released records), enabling per-program-type evaluation breakdowns; rank candidates that fail separation margins are dropped rather than emitted as ambiguous expressions. *(iii) Stratified:* candidates are distributed across target families within each query-type bucket using largest-remainder allocation, ensuring that no single family dominates within a bucket. T1 additionally separates multi-target queries into core and challenge sub-buckets to balance simpler family-all queries against more discriminative top-K and side-selection queries.

Selection and verification. For each final-pool image, the generator enumerates all template candidates and runs automatic checks for target-count consistency, eligibility-flag agreement, and absent-family validity. Selection follows the budget reported above, with multi-target candidates ranked by quality and density preferences. A stratified subsample across $\langle \text{source}, \text{query_type} \rangle$ cells is drawn for manual inspection; queries failing manual review are excluded from the released set. Table A6 lists the complete template library used in the released benchmark.

Table A6: **AgroVG query template library.** Each program type defines a triggering condition and a target-set resolution rule. The *Group* column gives a reader-friendly aggregation used in diagnostic analyses, while *Program type* is the exact identifier stored in the released query records.

Regime	Group	Program type	Tasks	Trigger condition	Target resolution
Single	Unique	`single_fine_unique`	T1, T2	Fine class has exactly one instance.	The unique instance.
Single	Unique	`single_family_unique`	T1, T2	Family has exactly one queryable instance.	The unique instance.
Single	Spatial rank	`single_rank_leftmost`	T1, T2	Family has $\geq$ 2 queryable instances; T1 requires separation.	Instance with smallest bbox-center x.
Single	Spatial rank	`single_rank_rightmost`	T1, T2	Family has $\geq$ 2 queryable instances; T1 requires separation.	Instance with largest bbox-center x.
Single	Spatial rank	`single_rank_topmost`	T1, T2	Family has $\geq$ 2 queryable instances; T1 requires separation.	Instance with smallest bbox-center y.
Single	Spatial rank	`single_rank_bottommost`	T1, T2	Family has $\geq$ 2 queryable instances; T1 requires separation.	Instance with largest bbox-center y.
Single	Size rank	`single_rank_largest`	T1, T2	Family has ≥ 2 queryable instances; T1 requires area separation.	Largest bbox area for T1; largest mask area for T2.
Single	Size rank	`single_rank_smallest`	T2	Family has ≥ 2 queryable instances.	Instance with smallest mask area.
Single	Relation	`single_relation_family`	T1	Crop and weed coexist with a unique anchor relation.	The unique target left/right/above/below the anchor.
Multi	All	`multi_family_all`	T1, T2	Family has ≥ 2 queryable instances and passes the count guardrail.	All queryable instances of that family.
Multi	Top-K	`multi_family_topk_leftmost`	T1, T2	Family has a separated top-K set.	Top-K by smallest bbox-center x.
Multi	Top-K	`multi_family_topk_largest`	T1, T2	Family has a separated top-K set.	Top-K by largest area.
Multi	Side subset	`multi_family_side_left`	T1	Valid left-side subset with at least two targets.	All queryable instances in the left-side subset.
Multi	Side subset	`multi_family_side_right`	T1	Valid right-side subset with at least two targets.	All queryable instances in the right-side subset.
Multi	Fine-class all	`multi_fine_all`	T1	Fine class has ≥ 2 queryable instances.	All queryable instances of that fine class.
Multi	Fine-class top-K	`multi_fine_topk_leftmost`	T1	Fine class has a separated top-K set.	Top-K by smallest bbox-center x within the fine class.
Multi	Fine-class top-K	`multi_fine_topk_largest`	T1	Fine class has a separated top-K set.	Top-K by bbox area within the fine class.
Multi	Fine-class side	`multi_fine_side_right`	T1	Realized fine-class right-side subset.	All queryable fine-class instances in the right-side subset.
Empty	Absent family	`empty_family_negative`	T1	Target family is absent from queryable instances.	Empty set.
Empty	Absent family	`empty_family_absent`	T2	Target family is absent from queryable instances.	Empty set.

Per-source composition. The $80\%/20\%$ dev/test split ratio is preserved at the per-source level (Table A7), ensuring that source-level evaluation breakdowns remain meaningful on the test set.

Table A7: **Per-source dev/test composition.** For each source, we report the number of image-query pairs in the dev and test splits and the total. The realized dev fraction stays close to the target 80% across all sources, reflecting the source-stratified split policy in Appendix B.3.

Task	Source	Dev	Test	Total	Dev frac.
T1	ACFR	509	127	636	80.0%
T1	CropAndWeed	505	127	632	79.9%
T1	Dumortier	377	94	471	80.0%
T1	GWHD 2021	889	202	1,091	81.5%
T1	IP102	1,004	251	1,255	80.0%
T1	MinneApple	480	118	598	80.3%
T1	OAM-TCD	83	21	104	79.8%
T1	PhenoBench	422	106	528	79.9%
T1	PlantSeg	969	242	1,211	80.0%
T1 subtotal		**5,238**	**1,288**	**6,526**	**80.3%**
T2	ACFR	67	17	84	79.8%
T2	CropAndWeed	354	89	443	79.9%
T2	MegaFruits-blueberry	443	111	554	80.0%
T2	MegaFruits-peach	468	117	585	80.0%
T2	PhenoBench	586	146	732	80.1%
T2	PlantSeg	918	229	1,147	80.0%
T2 subtotal		**2,836**	**709**	**3,545**	**80.0%**
AgroVG total		**8,074**	**1,997**	**10,071**	**80.2%**

The joint source–target-family distribution (Table A8) exhibits a diagonal-plus-off-diagonal structure: diagonal cells carry positive queries grounded in the source's primary family, while off-diagonal cells carry target-absent queries derived from verified family absence rather than synthetic mismatches. This structure enables existence-aware abstention evaluation without cross-source query mixing.

Table A8: **Per-source × per-target-family image-query pair counts.** Diagonal cells (positive queries grounded in the source's primary family) carry the bulk of each source's contribution; off-diagonal cells correspond to target-absent queries asking the model to abstain on families absent from the image. Empty cells are zero.

Task	Source	Crop	Disease	Fruit	Pest	Tree	Weed	Wheat head	Total
T1	ACFR	9	5	**613**	9				636
T1	CropAndWeed	**219**	35	29	21		**328**		632
T1	Dumortier		23	25	19	**404**			471
T1	GWHD 2021		66	45	71			**909**	1,091
T1	IP102	64	52	72	**1,067**				1,255
T1	MinneApple	57	48	**435**	58				598
T1	OAM-TCD		6	7	6	**85**			104
T1	PhenoBench	**275**	32	27	30		**164**		528
T1	PlantSeg	73	**1,030**	61	47				1,211
T1 total		697	1,297	1,314	1,328	489	492	909	**6,526**
T2	ACFR	5	5	**71**			3		84
T2	CropAndWeed	**186**	14	33			**210**		443
T2	MegaFruits-blueberry	24	38	**450**			42		554
T2	MegaFruits-peach	28	41	**482**			34		585
T2	PhenoBench	**499**	34	28			**171**		732
T2	PlantSeg	75	**944**	73			55		1,147
T2 total		817	1,076	1,137	—	—	515	—	**3,545**

T2 columns for pest, tree, and wheat head are empty by construction: the AgroVG-T2 global target family space is {crop, weed, fruit, disease} (Appendix B.4). Bold cells indicate the source's primary family or families.

Per-program-type query counts. Table A9 reports the realized query count per program type. T1 instantiates 18 program types covering single-target uniqueness and rank patterns, multi-target family-all and top-K patterns at both family and fine-class granularity, side-region selection, family-relation queries, and target-absent negatives. T2 instantiates a more restricted library of 12 program types focused on family-level templates: rank-based queries on individual instances within a fine-grained class often have ambiguous mask-level semantics and were excluded during template design. T2 additionally introduces `single_rank_smallest`, which is absent from T1 because the smallest-by-bbox-area instance is often visually ambiguous, while a smallest-mask instance carries pixel-level evidence that anchors the reference unambiguously.

Table A9: **Per-program-type query counts.** For each task we list the realized number of queries emitted by each program type, broken down into dev and test splits. Program types are grouped by query type (single, multi, empty). The total counts add up to the per-task headline counts (6,526 for T1 and 3,545 for T2).

Task	Query type	Program type	Dev	Test	Total
T1	single	`single_family_unique`	212	43	255
T1	single	`single_fine_unique`	202	53	255
T1	single	`single_rank_leftmost`	382	73	455
T1	single	`single_rank_rightmost`	338	75	413
T1	single	`single_rank_topmost`	371	86	457
T1	single	`single_rank_bottommost`	360	90	450
T1	single	`single_rank_largest`	685	196	881
T1	single	`single_relation_family`	24	5	29
T1	multi	`multi_family_all`	402	96	498
T1	multi	`multi_family_side_left`	63	15	78
T1	multi	`multi_family_side_right`	75	15	90
T1	multi	`multi_family_topk_leftmost`	664	172	836
T1	multi	`multi_family_topk_largest`	536	132	668
T1	multi	`multi_fine_all`	58	18	76
T1	multi	`multi_fine_side_right`	1	0	1
T1	multi	`multi_fine_topk_leftmost`	12	4	16
T1	multi	`multi_fine_topk_largest`	55	16	71
T1	empty	`empty_family_negative`	798	199	997
T1 subtotal			**5,238**	**1,288**	**6,526**
T2	single	`single_family_unique`	85	20	105
T2	single	`single_fine_unique`	81	32	113
T2	single	`single_rank_leftmost`	204	56	260
T2	single	`single_rank_rightmost`	211	55	266
T2	single	`single_rank_topmost`	217	40	257
T2	single	`single_rank_bottommost`	231	48	279
T2	single	`single_rank_largest`	217	45	262
T2	single	`single_rank_smallest`	171	59	230
T2	multi	`multi_family_all`	358	90	448
T2	multi	`multi_family_topk_leftmost`	306	71	377
T2	multi	`multi_family_topk_largest`	329	87	416
T2	empty	`empty_family_absent`	426	106	532
T2 subtotal			**2,836**	**709**	**3,545**

Limitations of dataset construction. The current release has two construction limitations. First, T2 is restricted to three of the six target families (crop/weed, fruit, plant disease), since pest, wheat-head, and tree-canopy sources do not yet provide audited instance-level pixel evidence under our protocol. Second, template-based query generation gives exact target-set traceability and deterministic verification but limits the linguistic diversity of the released queries, which do not cover many pragmatic references, dialogue contexts, or long contextual descriptions. The separate human-paraphrase study in Appendix D.4 tests wording sensitivity on a small T1 subset; it does not expand the linguistic coverage of the released benchmark. Future versions will incorporate additional mask-capable sources and human-rewritten or LLM-augmented expressions while preserving instance-level target-set verification.

C Evaluation Setup and Reproducibility

This section provides reproducibility-oriented details that complement the evaluation protocol described in Sec. 4: full metric definitions and edge-case conventions (Sec. C.1), model interfaces, prompts, and output parsing rules (Sec. C.2), and the compute environment and released artifacts that support score recomputation (Sec. C.3).

C.1 Metric Definitions and Edge Cases

Throughout this subsection, $\mathcal{Q}^+$ denotes positive queries whose ground-truth target set is non-empty, and $\mathcal{Q}^0$ denotes target-absent queries whose correct output is empty. For T2, mask-quality metrics are reported on $\mathcal{Q}^+$ and abstention metrics on $\mathcal{Q}^0$ unless explicitly labeled as all-query diagnostics.

T1 box-set matching. For a T1 query q, let $\mathcal{B}_q^*$ be the ground-truth box set and $\hat{\mathcal{B}}_q$ be the predicted box set after output normalization. At an IoU threshold $\tau \in \{0.50, 0.75\}$, we form a thresholded bipartite graph between ground-truth and predicted boxes:

$$E_\tau(q) = \{(b_i^*, \hat{b}_j) : \mathrm{IoU}(b_i^*, \hat{b}_j) \geq \tau,\ b_i^* \in \mathcal{B}_q^*,\ \hat{b}_j \in \hat{\mathcal{B}}_q\}.$$

We then compute a maximum-cardinality matching $M_\tau(q) \subseteq E_\tau(q)$. This objective counts how many instances are correctly localized at threshold τ, rather than maximizing the sum of IoU values among matched pairs; the choice reflects AgroVG's set-prediction formulation, where each ground-truth instance should be claimed by at most one prediction. Matched pairs are counted as true positives, unmatched predictions as false positives, and unmatched ground-truth boxes as false negatives:

$$\mathrm{TP}_\tau(q) = |M_\tau(q)|, \quad \mathrm{FP}_\tau(q) = |\hat{\mathcal{B}}_q| - |M_\tau(q)|, \quad \mathrm{FN}_\tau(q) = |\mathcal{B}_q^*| - |M_\tau(q)|.$$

The query-level Set-F_1 is the harmonic mean of the corresponding precision and recall:

$$F_{1,\tau}(q) = \frac{2\mathrm{TP}_\tau(q)}{2\mathrm{TP}_\tau(q) + \mathrm{FP}_\tau(q) + \mathrm{FN}_\tau(q)}.$$

Two boundary conventions handle empty-set cases consistently with the abstention semantics: $F_{1,\tau}(q) = 1$ when both $\mathcal{B}_q^*$ and $\hat{\mathcal{B}}_q$ are empty, and $F_{1,\tau}(q) = 0$ when $\mathcal{B}_q^*$ is empty but $\hat{\mathcal{B}}_q$ is not.

T1 aggregation, single-target accuracy, and abstention. For any query subset $\mathcal{S} \subseteq \mathcal{Q}$, macro Set-$F_1$ averages query-level scores, while micro Set-F_1 first sums true positives, false positives, and false negatives over $\mathcal{S}$ and then computes a single F_1:

$$\text{Macro-}F_{1,\tau}(\mathcal{S}) = \frac{1}{|\mathcal{S}|} \sum_{q \in \mathcal{S}} F_{1,\tau}(q),$$

$$\text{Micro-}F_{1,\tau}(\mathcal{S}) = \frac{2\sum_{q \in \mathcal{S}} \mathrm{TP}_\tau(q)}{2\sum_{q \in \mathcal{S}} \mathrm{TP}_\tau(q) + \sum_{q \in \mathcal{S}} \mathrm{FP}_\tau(q) + \sum_{q \in \mathcal{S}} \mathrm{FN}_\tau(q)}.$$

Macro Set-F_1 weights each query equally and is reported as the headline T1 metric. Micro Set-F_1 is computed by the released evaluator and included in the released summary files as a complementary box-level diagnostic, where high-cardinality queries contribute more strongly through their instance counts.

For single-target queries, we additionally report a strict accuracy that penalizes both cardinality errors and localization errors. A single-target query is counted as correct only when the model predicts exactly one valid box and that box matches the unique ground-truth box at threshold τ:

$$\text{S-Acc}_\tau = \frac{1}{|\mathcal{Q}_{\text{single}}|} \sum_{q \in \mathcal{Q}_{\text{single}}} \mathbb{1}\left[|\hat{\mathcal{B}}_q| = 1 \ \wedge \max_{b^* \in \mathcal{B}_q^*, \hat{b} \in \hat{\mathcal{B}}_q} \mathrm{IoU}(b^*, \hat{b}) \geq \tau\right].$$

For target-absent queries, the correct prediction is the empty set. Empty Accuracy is the fraction of target-absent queries with $|\hat{\mathcal{B}}_q| = 0$:

$$\text{E-Acc} = \frac{1}{|\mathcal{Q}^0|} \sum_{q \in \mathcal{Q}^0} \mathbb{1}\left[|\hat{\mathcal{B}}_q| = 0\right].$$

Empty Accuracy is reported separately rather than aggregated with positive-query localization diagnostics, because it measures existence-aware abstention rather than box-set overlap on non-empty target sets.

T2 query-level mask construction. For T2, evaluation operates on query-level binary masks rather than per-instance mask matching. This design reflects the heterogeneity of segmentation predictors evaluated in this paper, including per-instance segmentation models, query-level promptable text-to-mask models, and two-stage box-to-mask pipelines. Let $L_q \in \mathbb{Z}^{H\times W}$ be the image-level instance map and let $\mathcal{I}_q$ be the set of target instance IDs associated with query q. The ground-truth query mask is the union of all target instances:

$$M_q^*(i,j) = \mathbb{1}[L_q(i,j) \in \mathcal{I}_q].$$

For target-absent queries, $\mathcal{I}_q = \emptyset$ and M_q^* is the all-zero mask. The predicted mask $\hat{M}_q$ is binarized using the model's documented foreground convention and resized to the ground-truth resolution by nearest-neighbor interpolation when needed.

T2 mask-quality metrics on positive queries. For each query, intersection I_q, union U_q, ground-truth area A_q^*, and predicted area $\hat{A}_q$ are computed pixel-wise:

$$I_q = \sum_{i,j} M_q^*(i,j)\hat{M}_q(i,j), \quad U_q = \sum_{i,j} \mathbb{1}[M_q^*(i,j) \vee \hat{M}_q(i,j)],$$

$$A_q^* = \sum_{i,j} M_q^*(i,j), \qquad \hat{A}_q = \sum_{i,j} \hat{M}_q(i,j).$$

Per-query IoU and Dice are defined as

$$\mathrm{IoU}_q = \begin{cases} I_q/U_q, & U_q > 0, \\ 1, & U_q = 0, \end{cases} \qquad \mathrm{Dice}_q = \begin{cases} 2I_q/(A_q^* + \hat{A}_q), & A_q^* + \hat{A}_q > 0, \\ 1, & A_q^* + \hat{A}_q = 0. \end{cases}$$

The value 1 in the empty-empty case corresponds to correct abstention and is used only for explicitly labeled all-query diagnostic summaries. The main T2 mask-quality metrics are computed on positive queries only:

$$\mathrm{mIoU}^+ = \frac{1}{|\mathcal{Q}^+|}\sum_{q\in\mathcal{Q}^+} \mathrm{IoU}_q, \qquad \mathrm{mDice}^+ = \frac{1}{|\mathcal{Q}^+|}\sum_{q\in\mathcal{Q}^+} \mathrm{Dice}_q,$$

$$\mathrm{cIoU}^+ = \frac{\sum_{q\in\mathcal{Q}^+} I_q}{\sum_{q\in\mathcal{Q}^+} U_q}, \qquad \mathrm{cDice}^+ = \frac{2\sum_{q\in\mathcal{Q}^+} I_q}{\sum_{q\in\mathcal{Q}^+} A_q^* + \sum_{q\in\mathcal{Q}^+} \hat{A}_q}.$$

Thresholded mask success is reported as

$$\mathrm{IoU@}\tau^+ = \frac{1}{|\mathcal{Q}^+|}\sum_{q\in\mathcal{Q}^+} \mathbb{1}[\mathrm{IoU}_q \geq \tau], \qquad \tau \in \{0.50, 0.75\}.$$

The query-averaged metrics treat each query equally and reflect typical query difficulty; the cumulative metrics aggregate intersections, unions, and areas before taking the ratio and are therefore more sensitive to large masks. We use thresholded success rates rather than AP-style metrics because the evaluated models either do not expose confidence scores or expose scores with incompatible calibration.

T2 target-absent metrics. For target-absent queries, evaluation measures abstention rather than foreground overlap. With empty-area tolerance $\delta_e = 0$, Empty Accuracy is the fraction of target-absent queries with $\hat{A}_q \leq \delta_e$:

$$\text{E-Acc} = \frac{1}{|\mathcal{Q}^0|}\sum_{q\in\mathcal{Q}^0} \mathbb{1}[\hat{A}_q \leq \delta_e],$$

and the corresponding empty false-positive rate is

$$\text{Empty FPR} = 1 - \text{E-Acc}.$$

As in T1, target-absent behavior is reported as a stand-alone metric so that pixel-level overlap on positive queries and existence-aware abstention can be analyzed independently.

Family-macro reporting. Family-macro scores are computed by first evaluating the metric within each target family and then averaging over families with at least one valid query:

$$\mathrm{FamMacro}(m) = \frac{1}{|\mathcal{F}_{\mathrm{eval}}|} \sum_{f \in \mathcal{F}_{\mathrm{eval}}} m(\mathcal{Q}_f).$$

This prevents large sources or frequent target families from dominating the aggregate score and makes cross-family differences visible in the diagnostic breakdowns.

Malformed and missing outputs. All predictions pass through the official AgroVG output normalizer before scoring (Sec. C.2). Missing or unparsable T1 outputs are treated as empty box sets, while missing or unparsable T2 outputs are treated as all-zero masks; degenerate boxes, invalid coordinates, duplicate boxes, missing masks, and mask-shape mismatches are handled deterministically by the released evaluator. These conventions ensure that every query contributes exactly once to evaluation. However, failures normalized to empty predictions receive zero grounding credit on positive queries and may receive correct-empty credit on target-absent queries. We therefore report output-handling sensitivity separately from the primary scores (Appendix C.2).

C.2 Evaluated Models, Prompts, and Parsing

This subsection details the model interfaces, prompts, and output parsing rules used to produce the predictions evaluated in Sec. 5. The 26 task-specific configurations are summarized in Table A10. They span closed-source MLLMs, open-source VLMs, and specialized grounding systems evaluated through their native output interfaces or two-stage grounding-to-mask pipelines. Closed-source MLLMs and open-source VLMs are evaluated on T1 box grounding, while specialized systems cover T1 detectors, T2 segmentation MLLMs, T2 promptable text-to-mask models, and T2 two-stage pipelines. Florence-2-Large and OWLv2-Large appear in multiple task-specific configurations because they are evaluated through different output interfaces. We use a fixed interface-level prompting protocol across all configurations and select detector thresholds and prompt formats only on the development split.

Prompting protocol. Prompts are specified at the interface level rather than tuned per model. For instruction-following MLLMs and VLMs without a documented native grounding syntax, T1 queries use the following prompt:

```
You are solving an agricultural visual grounding task.
Given an image and a natural-language query, return all bounding boxes
for objects that match the query.

Query: {query}

Return JSON only:
{"boxes": [[xmin, ymin, xmax, ymax], ...]}

Coordinates must be normalized to [0,1] and use xyxy order.
If no object matches the query, return {"boxes": []}.
Include every matching object and only matching objects.
```

The prompt specifies the output container, coordinate convention, target-absent behavior, and cardinality rule. Models with documented native grounding syntax use their interface-specific formats: DeepSeek-VL2 uses `<|ref|>{query}<|/ref|>`; Florence-2-Large uses `<OPEN_VOCABULARY_DETECTION>{query}` for T1 and `<REFERRING_EXPRESSION_SEGMENTATION>{query}` for T2; T1 grounding detectors such as OWLv2-Large and Grounding DINO 1.5 receive the raw AgroVG query text without an instruction-style preamble. For T2, promptable and two-stage mask systems receive the raw segmentation query directly, while segmentation MLLMs that require an explicit segmentation trigger append the suffix "`Please output segmentation mask.`" to the query. The same prompt family is used for all queries within a model configuration; prompts are not retuned on the test split.

T1 output normalization. The T1 parser converts heterogeneous model outputs into a canonical list of absolute-pixel `xyxy` boxes. Structured JSON outputs are parsed directly when possible. For nested JSON, tagged box formats, or valid boxes embedded in free-form text, the parser applies documented recovery rules before conversion to the AgroVG box format. Coordinate conventions are normalized to absolute pixels, including pixel coordinates, $[0, 1]$ normalized coordinates, and $[0, 1000]$ coordinate formats used by some VLMs. Florence-2 outputs are processed by its official post-processor before conversion; DeepSeek-VL2 tagged boxes are converted from the model's normalized coordinate system to image pixels. All boxes are clipped to image bounds, and predictions with non-positive area, degenerate full-image extent, or exact duplication are removed before scoring.

T2 output normalization. The T2 parser converts each model output into one query-level binary mask saved as a grayscale PNG. Direct mask outputs are binarized and unioned via pixel-wise logical OR when multiple masks are returned. Polygon outputs are rasterized onto the original image grid and unioned. Segmentation-token models are decoded through their native mask decoders, and all `[SEG]`-derived masks are unioned. For two-stage pipelines, masks produced by SAM from each detector box are unioned into a single query-level prediction. If a model produces no usable mask, the prediction is represented as an all-zero mask saved at the canonical path, ensuring deterministic representation at evaluation time. Each prediction record stores the query ID and the mask path, allowing the official evaluator to recompute all T2 metrics without rerunning inference.

Detector thresholds and parser warnings. Detector thresholds, when required, are selected on the development split and held fixed for the test split. The official evaluator emits per-run warning files documenting missing predictions, malformed boxes, missing mask paths, mask-shape mismatches, and other normalization events. These warning logs are released alongside the predictions so that output-format failures can be audited separately from model-level grounding errors.

Output-handling sensitivity. We audit the frozen outputs for all $1{,}288$ T1 test queries, treating valid explicit empty sets as abstention rather than failure. A separate deterministic recovery procedure uses output syntax and documented coordinate conventions without access to ground truth; blank/API failures and degenerate geometry remain unrepaired. Table A11 reports changes relative to the unchanged primary evaluation policy. Overall Set-F_1 at IoU 0.50 changes by less than 0.0001 for nine configurations. The largest increase, 0.0371 for InternVL3.5-38B, comes from coordinate-scale inference rather than JSON parsing.

For Gemini 2.5 Pro, recovery increases positive-query Set-F_1 by 0.0423 while reducing Empty Accuracy by 0.1457, removing accidental empty-set credit on 29 of the 199 target-absent queries. These 29 cases account for 29 of its 75 primary empty-query successes. The only ordering reversal is between DeepSeek-VL2-small and Qwen3-VL-8B, whose recovered overall scores are 0.3272 and 0.3271. Thus output handling can affect positive grounding and apparent abstention in opposite directions. Primary scores remain unchanged, and recovery results are reported only as sensitivity diagnostics. For T2, 12 of 13 configurations provide evaluable masks for all 709 test queries; Florence-2-Large has one missing positive-query mask.

C.3 Compute and Reproducibility Artifacts

AgroVG evaluates existing models without training or fine-tuning, so the benchmark's compute footprint is dominated by inference. This subsection documents the compute environment, the released artifacts, and the two reproducibility modes that the release supports.

Compute environment. Open-source VLMs and specialized grounding systems are evaluated locally on NVIDIA A100 80GB GPUs using each model's official loader, checkpoint, and preprocessing pipeline. Closed-source MLLMs are queried through their official APIs. For every run, we record a manifest containing the checkpoint identifier or provider model ID, query date for API runs, code revision when available, precision and decoding settings, detector thresholds, prompt family, parser version, hardware configuration, and the raw output before parsing. We use deterministic or greedy decoding whenever supported, and detector thresholds selected on the development split are fixed before test-set evaluation. The local evaluation campaign for open-source and specialized

Table A10: **Evaluated model registry.** We report the 26 task-specific configurations evaluated on AgroVG, including API identifiers or public checkpoints, task assignment, and output interface. "Native" indicates that the model directly emits the evaluated output, while "Two-stage" indicates detector-based grounding followed by SAM mask generation.

Model	Params.	Task	API identifier / checkpoint	Interface
Closed-source MLLMs				
GPT-4o	–	T1	`gpt-4o-2024-08-06`	OpenAI API, box output
GPT-5.4	–	T1	`gpt-5.4`	OpenAI API, box output
Gemini 2.5 Pro	–	T1	`gemini-2.5-pro`	Google API, box output
Claude Sonnet 4.6	–	T1	`claude-sonnet-4-6`	Anthropic API, box output
Open-source VLMs				
DeepSeek-VL2-tiny	∼1B	T1	`deepseek-ai/deepseek-vl2-tiny`	Native / prompted box output
DeepSeek-VL2-small	∼2.8B	T1	`deepseek-ai/deepseek-vl2-small`	Native / prompted box output
InternVL3.5-8B	8.53B	T1	`OpenGVLab/InternVL3_5-8B`	Prompted box output
InternVL3.5-38B	38.4B	T1	`OpenGVLab/InternVL3_5-38B`	Prompted box output
Qwen3-VL-8B	9B	T1	`Qwen/Qwen3-VL-8B-Instruct`	Prompted box output
Qwen3-VL-32B	33B	T1	`Qwen/Qwen3-VL-32B-Instruct`	Prompted box output
Specialized T1 grounding detectors				
OWLv2-Large	0.4B	T1	`google/owlv2-large-patch14-ensemble`	Native box output
Florence-2-Large	0.77B	T1	`microsoft/Florence-2-large`	Native box output
Grounding DINO 1.5	–	T1	`IDEA-Research/Grounding-DINO-1.5-API`	Native box output
T2 segmentation MLLMs / reasoning-segmentation models				
LISA-7B	7B	T2	`xinlai/LISA-7B-v1`	Native mask output
LISA-13B	13B	T2	`xinlai/LISA-13B-llama2-v1`	Native mask output
GLaMM	7B	T2	`MBZUAI/groundingLMM-FullScope`	Native mask output
PixelLM-7B	7B	T2	`MaverickRen/PixelLM-7B`	Native mask output
PixelLM-13B	13B	T2	`MaverickRen/PixelLM-13B`	Native mask output
GSVA (force-seg)	13B	T2	`LeapLabTHU/GSVA-13B-ft`	Native mask output
PSALM	2B	T2	`EnmingZhang/PSALM`	Native mask output
T2 promptable / universal text-to-mask models				
X-Decoder	–	T2	`microsoft/X-Decoder`	Native mask output
SEEM	–	T2	`xdecoder/SEEM`	Native mask output
Florence-2-Large	0.77B	T2	`microsoft/Florence-2-large`	Native mask output
SAM 3	0.9B	T2	`facebook/sam3`	Native mask output
T2 two-stage box-to-mask pipelines				
Grounded-SAM	Comp.	T2	GroundingDINO-T + `facebook/sam-vit-huge`	Two-stage
OWLv2-Large + SAM	Comp.	T2	OWLv2-Large + `facebook/sam-vit-huge`	Two-stage

For API-based models, released run metadata records the provider identifier, query date, prompt family, decoding parameters, and raw response before parsing. For local models, run manifests record checkpoint, precision, decoding settings, thresholds where applicable, and parser version. "Comp." denotes a composite pipeline rather than a single parameterized model.

Table A11: T1 sensitivity to additional output normalization. Changes are relative to the unchanged primary scores. Pos-F_1 averages Set-F_1 over positive queries only; E-Acc evaluates target-absent queries. "Other eight" comprises GPT-4o, GPT-5.4, Claude Sonnet 4.6, DeepSeek-VL2-tiny, InternVL3.5-8B, Florence-2-Large, OWLv2-Large, and Grounding DINO 1.5.

Model	Δ All@.5	Δ Pos-F_1@.5	Δ E-Acc
Gemini 2.5 Pro	$+0.0133$	$+0.0423$	-0.1457
DeepSeek-VL2-small	$+0.0081$	$+0.0113$	-0.0101
InternVL3.5-38B	$+0.0371$	$+0.0449$	-0.0050
Qwen3-VL-8B	$\lvert\Delta\rvert < 0.0001$	$+0.0009$	-0.0050
Qwen3-VL-32B	$+0.0035$	$+0.0050$	-0.0050
Other eight	0.0000	0.0000	0.0000

configurations required approximately 320 A100 GPU-hours in total, covering inference, output normalization, and metric computation; closed-source API compute is borne by the providers and is not observable from the client.

Released evaluation artifacts. The AgroVG release contains all non-provider artifacts needed to audit and recompute the reported metrics, organized into three groups. *Data:* fixed dev/test split files, normalized image and instance metadata, and query JSONL files. *Protocol:* prompt templates, parsing scripts, and the official evaluation scripts. *Predictions:* per-model prediction JSONLs, saved T2 mask predictions, per-query result files, aggregate `summary.json` files, and parser-warning logs. Every released query maps to exactly one normalized prediction before scoring, so the official evaluator reproduces the paper tables from the released predictions without rerunning any model inference.

Reproducibility modes. The released artifacts support two reproducibility modes. *Score recomputation:* users run the official evaluator on the released predictions and recover the aggregate, per-source, per-family, per-query-type, and per-program-type metrics reported in the paper; this mode requires no GPU inference. *Full inference reproduction:* users obtain the original source images under their respective licenses, load the listed checkpoints or query the listed APIs, apply the released prompt and parsing protocol, and rerun the evaluator. Full reproduction is intended to be deterministic for locally evaluated models under the recorded decoding settings, software versions, and parsing rules; minor numerical differences may still arise from hardware or library changes. For closed-source APIs, exact reproduction may differ because provider-side model serving is outside the benchmark's control. The released raw responses, parsed predictions, model identifiers, and query dates define the auditable record for the reported API scores and allow the published tables to be recomputed even if a provider later changes the model behind a public endpoint.

D Additional Results and Diagnostics

This section analyzes frozen test predictions by source and query program (Appendices D.1–D.2), summarizes observed failure patterns (Appendix D.3), and quantifies uncertainty in model comparisons (Appendix D.4). Metrics follow Appendix C.1. Complete T1 query-type and T2 target-family breakdowns accompany the source package as CSV files in `supplementary_results/`.

D.1 T1 Diagnostic Results

We report macro Set-F_1 at IoU 0.50 from the frozen T1 predictions. Values in the following tables are percentages; the overall column in Table A12 reproduces the main T1 score up to rounding.

Source-level variation. Table A12 shows substantial variation across agricultural sources. These are descriptive comparisons across subsets with different viewpoints, target sizes, instance densities, and background clutter. Models with similar overall scores can have different source-level profiles.

Single-target scoring. Single-target Set-F_1 gives partial credit for predictions containing a matched target and extra boxes, whereas strict accuracy requires exactly one correct box. For example, OWLv2-Large obtains 19.7% single-target Set-F_1 but 8.1% strict accuracy; Grounding DINO 1.5 obtains 28.8% and 14.2%, respectively. The complete query-type breakdown is provided in `t1_query_type_results.csv`; multi-target and target-absent scores appear in Table 2.

Table A12: **T1 per-source macro Set-F_1 at IoU $\tau = 0.50$.** Each cell is the macro-averaged Set-F_1 over the model's per-query results within the corresponding source. The rightmost column reproduces the overall T1 score from Table 2 as a consistency check. All numbers are percentages.

Model	ACFR	CW	Dumort.	GWHD	IP102	Minne.	OAM-TCD	Pheno.	PlantSeg	All
Closed-source MLLMs										
GPT-4o	3.1	20.9	15.8	17.3	22.6	15.3	14.3	8.5	11.1	15.0
GPT-5.4	8.8	37.6	19.5	16.0	62.7	16.0	14.3	28.3	13.9	27.4
Gemini 2.5 Pro	31.0	18.4	11.5	7.4	28.0	13.5	4.8	35.5	10.3	18.5
Claude Sonnet 4.6	6.7	20.9	13.8	8.5	48.6	11.9	15.9	25.5	10.5	20.0
Open-source VLMs										
DeepSeek-VL2-tiny	5.4	23.3	17.0	17.8	37.7	41.0	14.3	15.7	14.7	22.3
DeepSeek-VL2-small	18.2	45.4	26.2	10.4	70.5	31.5	21.6	29.2	14.4	31.9
InternVL3.5-8B	62.3	38.4	34.8	18.3	71.4	39.9	20.2	38.2	17.2	39.6
InternVL3.5-38B	61.7	31.8	44.9	22.2	73.4	24.8	26.2	39.6	18.9	39.8
Qwen3-VL-8B	36.0	21.4	29.8	23.8	63.3	29.1	14.3	33.4	16.8	32.7
Qwen3-VL-32B	55.7	21.2	41.2	24.5	73.8	15.3	14.3	35.2	22.1	37.5
Specialized grounding detectors										
OWLv2-Large	28.8	46.8	16.7	18.0	51.5	34.4	29.7	28.7	10.5	29.5
Florence-2-Large	38.9	25.4	9.3	1.7	37.2	21.3	14.6	20.3	9.6	20.2
Grounding DINO 1.5	38.1	31.3	13.3	1.9	54.9	15.6	10.3	21.6	13.5	24.7

CW = CropAndWeed; Dumort. = Dumortier; Minne. = MinneApple; Pheno. = PhenoBench.

Query-program variation. Table A13 groups the 18 T1 programs into eight buckets. Most models score higher on unique-target queries than on spatial- or size-rank queries. The multi-target

buckets further distinguish retrieving all matching instances from selecting a top-K or spatially constrained subset. These breakdowns describe performance under different query requirements without attributing errors to a single underlying capability.

Table A13: **T1 per-program-bucket macro Set-F_1 at IoU $\tau = 0.50$.** The 18 T1 program types (Appendix B.4) are aggregated into 8 program-type buckets. Each cell is the count-weighted average of the constituent program-type macro F_1 values. All numbers are percentages.

Model	S-uniq.	S-rank-sp.	S-rank-size	S-rel.	M-all	M-topk	M-side	Empty
GPT-4o	15.6	1.2	4.6	0.0	3.8	0.8	0.0	79.9
GPT-5.4	67.7	15.4	13.8	0.0	32.7	10.8	17.6	66.8
Gemini 2.5 Pro	15.6	16.4	10.7	0.0	27.9	11.2	19.9	37.7
Claude Sonnet 4.6	59.4	8.3	10.7	0.0	24.1	7.3	6.3	49.7
DeepSeek-VL2-tiny	6.2	12.3	9.2	0.0	3.7	5.1	26.8	97.5
DeepSeek-VL2-small	66.9	27.7	17.5	60.0	50.7	14.3	18.7	55.3
InternVL3.5-8B	77.6	36.2	17.9	40.0	55.7	28.7	30.6	58.3
InternVL3.5-38B	79.1	32.4	20.1	33.3	47.0	23.5	33.1	75.9
Qwen3-VL-8B	65.9	24.1	10.6	20.0	28.7	13.6	15.0	88.9
Qwen3-VL-32B	70.7	29.6	17.3	20.0	42.7	19.9	32.2	80.9
OWLv2-Large	62.4	13.5	9.1	17.4	24.3	13.3	10.5	92.5
Florence-2-Large	63.5	18.5	21.0	53.3	38.8	15.1	6.9	0.0
Grounding DINO 1.5	77.2	20.2	19.3	33.3	54.6	21.2	17.9	1.5

S-uniq. = Single-unique (`single_fine_unique` + `single_family_unique`); S-rank-sp. = Single-rank-spatial (4 spatial directions); S-rank-size = `single_rank_largest`; S-rel. = `single_relation_family`; M-all = `multi_family_all` + `multi_fine_all`; M-topk = 4 topk templates; M-side = side-selection templates realized in the test split; Empty = `empty_family_negative` (Empty Accuracy reported in this column).

D.2 T2 Diagnostic Results

T2 diagnostics distinguish positive-query mask quality from target-absent abstention, following Appendix C.1. Unless explicitly labeled all-query, mask-quality metrics use $\mathcal{Q}^+$ and Empty Accuracy uses $\mathcal{Q}^0$.

Source-level diagnostics. Table A14 reports auxiliary all-query mIoU by source, combining positive-query segmentation and target-absent abstention. It should therefore be interpreted alongside the separate positive-query and empty-query results rather than as a pure measure of segmentation quality. The analogous target-family breakdown is available in `t2_family_all_query_miou.csv`.

Table A14: **T2 per-source all-query mIoU diagnostic.** Each cell averages per-query IoU over all T2 test queries from the corresponding source, including positive and target-absent queries under the empty-mask convention in Appendix C.1. The table therefore reflects both positive-query mask quality and target-absent abstention. All numbers are percentages.

Model	Pheno.	CW	Blueb.	Peach	ACFR	PlantSeg	All
Reasoning-segmentation MLLMs							
LISA-7B	32.9	32.9	28.6	31.7	18.5	17.4	26.7
LISA-13B	30.1	37.4	30.6	42.8	12.6	15.8	28.1
GLaMM	27.0	26.1	26.7	24.2	14.3	17.2	22.9
PixelLM-7B	24.0	28.5	32.5	41.7	26.4	15.7	26.2
PixelLM-13B	20.2	29.1	32.3	41.1	25.4	14.6	25.0
GSVA (force-seg)	20.9	17.3	15.6	23.6	6.3	16.7	18.4
PSALM	29.4	23.5	29.2	41.1	33.6	13.0	25.3
Promptable / universal text-to-mask models							
X-Decoder	10.5	18.6	18.4	26.7	31.1	17.8	18.3
SEEM	20.1	23.3	19.1	29.5	25.8	14.6	20.3
Florence-2-Large	10.5	17.4	28.5	30.3	20.5	14.3	18.9
SAM 3	12.7	9.6	32.6	42.8	37.3	21.5	23.8
Two-stage box-to-mask pipelines							
Grounded-SAM	23.4	31.4	17.6	24.0	11.9	16.1	21.0
OWLv2-Large + SAM	36.8	9.2	16.6	18.3	11.8	13.6	19.0

Pheno. = PhenoBench; CW = CropAndWeed; Blueb. = MegaFruits-blueberry; Peach = MegaFruits-peach; ACFR = ACFR-apples.

Query-regime diagnostics. Table A15 reports positive-query mIoU and IoU@0.50 for single-target and multi-target queries, with Empty Accuracy in a separate column. This decomposition distinguishes mask coverage on present targets from empty-output behavior.

Table A15: **T2 per-query-type breakdown.** For single-target and multi-target positive queries, we report mIoU and IoU@0.50. For target-absent queries, we report Empty Accuracy. Unlike the all-query diagnostic tables, this table keeps positive mask quality and abstention behavior separate. All numbers are percentages.

	Single		Multi		Empty
Model	mIoU	IoU@.5	mIoU	IoU@.5	E-Acc
LISA-7B	23.2	21.4	41.8	44.4	2.8
LISA-13B	25.0	21.7	41.5	44.0	7.5
GLaMM	25.5	22.5	27.4	21.4	3.8
PixelLM-7B	28.7	28.2	33.8	29.8	0.0
PixelLM-13B	26.4	25.1	32.7	29.4	1.9
GSVA (force-seg)	16.2	11.5	29.3	27.4	0.0
PSALM	26.4	26.5	34.7	33.5	0.0
X-Decoder	11.8	11.3	18.8	17.3	38.7
SEEM	14.0	13.5	25.4	25.8	29.2
Florence-2-Large	21.4	18.9	14.9	6.0	19.8
SAM 3	9.3	6.2	14.9	15.3	93.4
Grounded-SAM	21.4	21.7	28.9	25.4	0.9
OWLv2-Large + SAM	8.6	8.2	13.0	10.5	67.9

Query-program diagnostics. Table A16 groups the 12 T2 programs into six buckets. Rank-based queries require selecting an instance by spatial position or mask size, while multi-target queries evaluate coverage of a union mask. Their scores describe sensitivity to these query requirements; they do not isolate referent-selection errors from mask-boundary errors.

Table A16: **T2 per-program-bucket diagnostics.** The 12 T2 program types are aggregated into six diagnostic buckets. Non-empty program buckets report count-weighted mIoU over positive queries, while the Empty bucket reports Empty Accuracy on target-absent queries. All numbers are percentages.

Model	S-uniq.	S-rank-sp.	S-rank-size	M-all	M-topk	Empty
LISA-7B	46.0	21.0	16.0	54.2	34.8	2.8
LISA-13B	45.3	23.6	17.5	50.5	36.3	7.5
GLaMM	48.7	23.3	18.2	29.5	26.2	3.8
PixelLM-7B	42.1	30.1	19.4	41.1	29.6	0.0
PixelLM-13B	36.8	27.8	18.8	38.6	29.4	1.9
GSVA (force-seg)	31.8	13.8	12.9	29.8	29.0	0.0
PSALM	38.9	26.0	20.8	46.3	28.2	0.0
X-Decoder	17.4	10.1	12.2	30.3	12.3	38.7
SEEM	22.4	11.1	15.3	38.3	18.1	29.2
Florence-2-Large	22.4	24.7	14.7	15.4	14.6	19.8
SAM 3	9.4	9.3	9.3	27.0	8.0	93.4
Grounded-SAM	50.0	15.7	18.1	31.7	27.3	0.9
OWLv2-Large + SAM	12.8	8.0	7.8	11.3	13.9	67.9

S-uniq. = Single-unique; S-rank-sp. = Single-rank-spatial (4 spatial directions); S-rank-size = `single_rank_largest` + `single_rank_smallest`; M-all = `multi_family_all`; M-topk = 2 topk templates; Empty = Empty Accuracy on `empty_family_absent`.

Source-specific abstention. Table A17 reports Empty Accuracy by source. SAM 3 scores highly across sources, while LISA and PixelLM variants usually produce nonempty masks on queries labeled target-absent. X-Decoder and SEEM show larger source variation, with higher empty-query scores on PhenoBench and CropAndWeed than on the MegaFruits subsets.

Table A17: **T2 per-source Empty Accuracy on target-absent queries.** Each cell is the fraction of target-absent queries within the source for which the model emits no foreground mask, i.e., $\hat{A}_q \leq \delta_e = 0$. The Overall column reports Empty Accuracy across all target-absent T2 queries. All numbers are percentages.

Model	Pheno.	CW	Blueb.	Peach	ACFR	PlantSeg	All
LISA-7B	21.4	0.0	0.0	0.0	0.0	0.0	2.8
LISA-13B	0.0	0.0	10.5	33.3	0.0	0.0	7.5
GLaMM	14.3	0.0	0.0	11.1	0.0	0.0	3.8
PixelLM-7B	0.0	0.0	0.0	0.0	0.0	0.0	0.0
PixelLM-13B	0.0	0.0	0.0	5.6	0.0	2.1	1.9
GSVA (force-seg)	0.0	0.0	0.0	0.0	0.0	0.0	0.0
PSALM	0.0	0.0	0.0	0.0	0.0	0.0	0.0
X-Decoder	64.3	80.0	10.5	5.6	50.0	50.0	38.7
SEEM	71.4	100.0	5.3	11.1	50.0	25.0	29.2
Florence-2-Large	28.6	100.0	0.0	5.6	0.0	22.9	19.8
SAM 3	100.0	100.0	78.9	94.4	100.0	95.8	93.4
Grounded-SAM	0.0	20.0	0.0	0.0	0.0	0.0	0.9
OWLv2-Large + SAM	85.7	100.0	68.4	72.2	100.0	56.2	67.9

D.3 Qualitative Failure Patterns and Case Studies

We complement the aggregate diagnostics with five illustrative cases from the `test` split: two T1 predictions by InternVL3.5-38B and three T2 predictions by LISA-13B (Figs. A2–A6). These examples show box-coverage, target-selection, and mask-boundary behavior on weed, tree-canopy, fruit, and plant-disease targets. They illustrate individual predictions rather than estimating failure frequencies or establishing a performance upper bound.

T1: box coverage and overlapping predictions. In the weed example, the predicted count matches the two annotated targets, but the boxes are smaller and shifted relative to the references (Fig. A2). In the tree-canopy example, two large, overlapping predictions contrast with references that delineate the targets separately (Fig. A3). These cases illustrate why target count alone is insufficient to assess box-set grounding. Green boxes denote references and red boxes denote predictions.

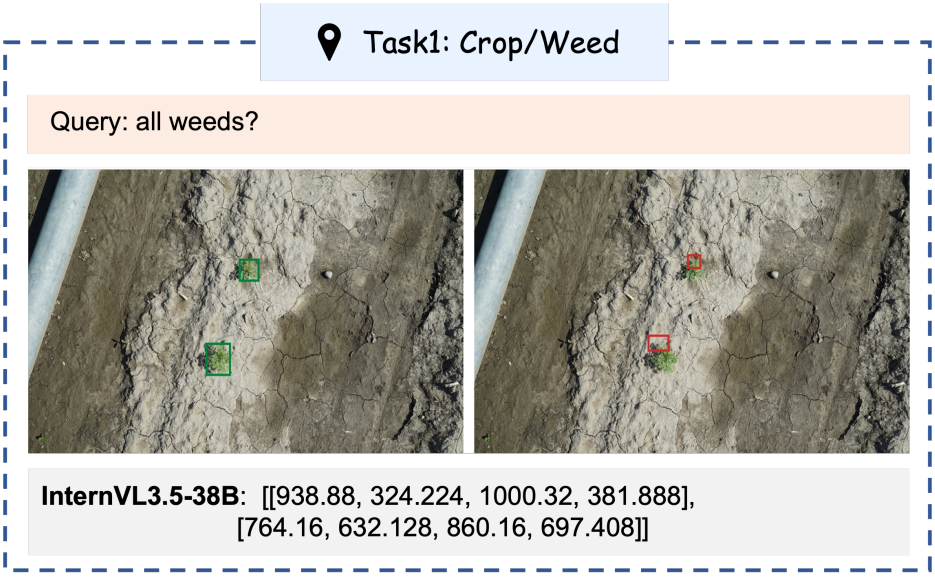


Figure A2: **T1 weed localization for "all weeds?"** InternVL3.5-38B returns two boxes, matching the number of annotated targets, but both predictions are smaller and shifted upward relative to the reference boxes. Green: reference boxes; red: predictions.

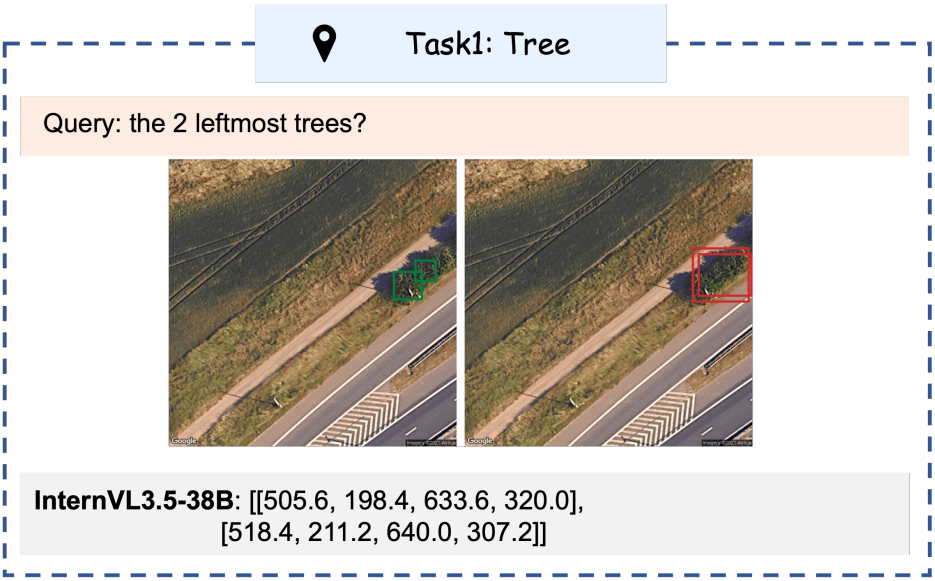


Figure A3: **T1 tree-canopy localization for "the 2 leftmost trees?"** InternVL3.5-38B returns two large, strongly overlapping boxes over neighboring canopies, whereas the reference boxes delineate the two targets separately. Green: reference boxes; red: predictions.

T2: target selection and mask boundaries. For a query requesting the bottommost fruit, LISA-13B marks two fruit regions instead of isolating one (Fig. A4). In the all-fruit example, the prediction extends into adjacent foliage and background (Fig. A5). The disease example shows a connected prediction over an irregular lesion region (Fig. A6). These panels display the input image on the left and the predicted foreground in red on the right; ground-truth masks are not shown. Boundary accuracy is measured by the quantitative mask metrics rather than inferred from these visualizations alone.

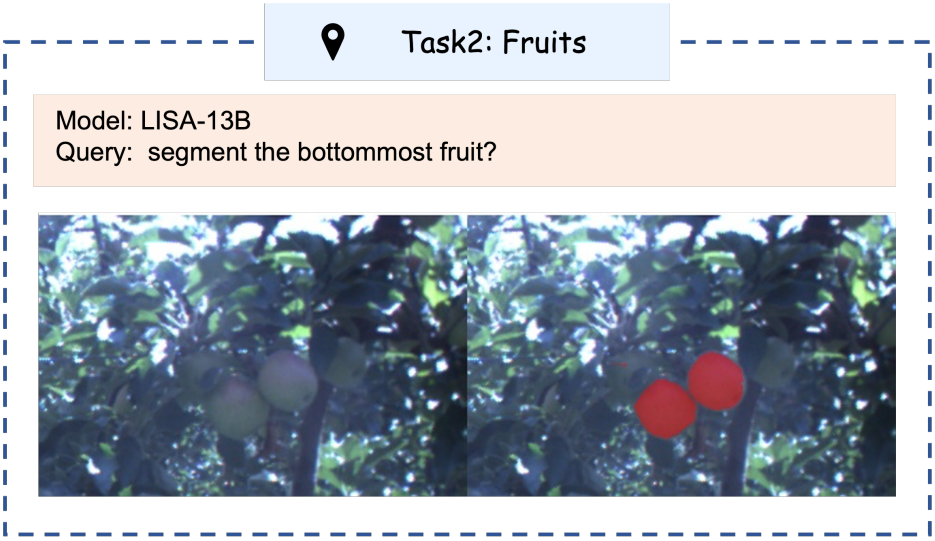


Figure A4: **T2 single-target fruit query.** For "segment the bottommost fruit?", LISA-13B predicts foreground on two fruit regions, rather than only the requested bottommost instance. Left: input image; right: predicted mask in red.

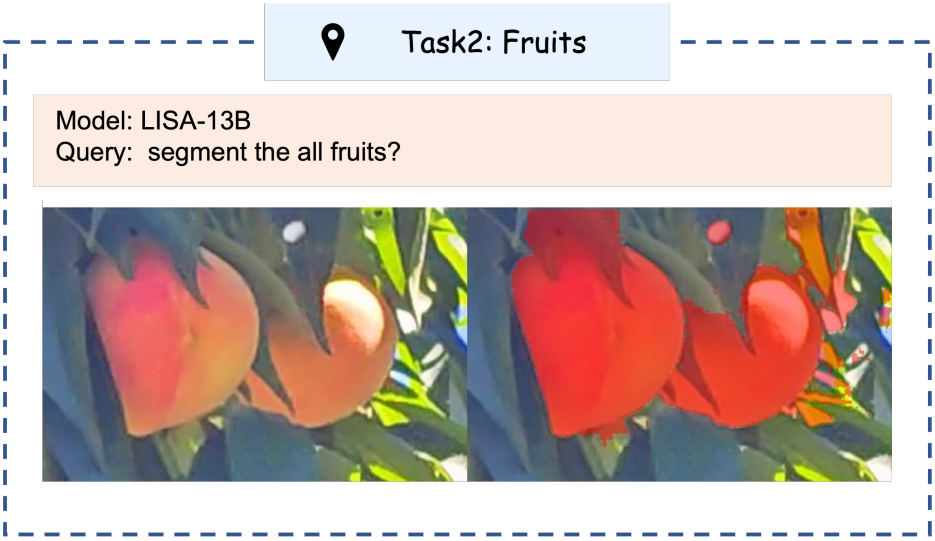


Figure A5: **T2 multi-target fruit query.** LISA-13B covers the visible fruit regions but also places foreground on adjacent foliage and background. Left: input image; right: predicted mask in red.

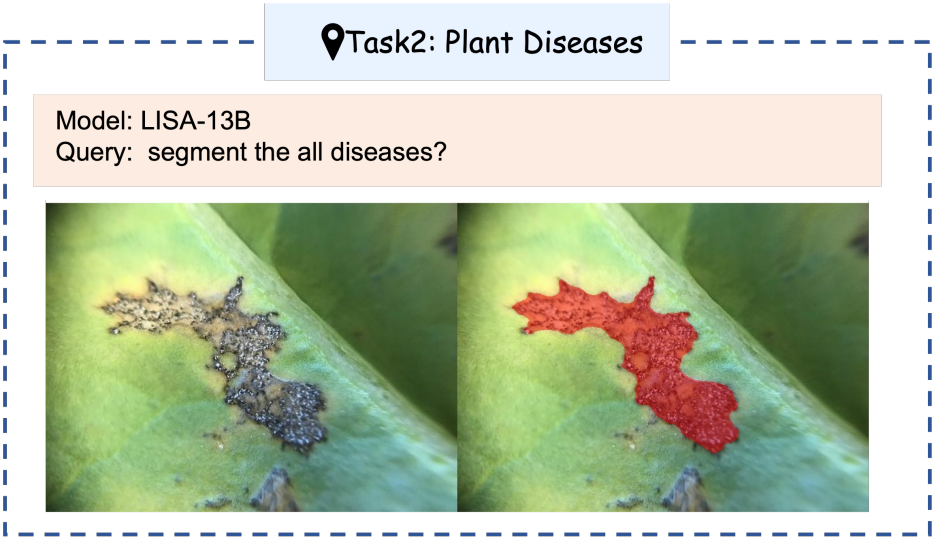


Figure A6: **T2 plant-disease query.** LISA-13B produces a connected mask over the main irregular lesion region, bridging some of its visible internal gaps. Left: input image; right: predicted mask in red.

Abstention. The illustrated cases contain positive targets. Target-absent behavior is instead summarized by the Empty Accuracy breakdowns in Appendices D.1 and D.2. Some models produce boxes or foreground masks on queries labeled target-absent, while others return empty predictions on positive queries. Empty outputs can also result from interface failures (Appendix C.2); these summaries do not by themselves establish the cause of an empty or nonempty prediction.

D.4 Uncertainty in Model Comparisons

We estimate uncertainty in model differences using 10,000 paired image-level bootstrap resamples, stratified by source and query type. Each resample uses the same images across all 13 configurations within a task, with scores computed from frozen test predictions. We compare adjacent configurations ordered by overall Set-F_1 at IoU 0.50 for T1 and positive-query mean IoU (All-mIoU) for T2.

Pointwise 95% confidence intervals include zero for four of the twelve adjacent T1 comparisons and seven of the twelve adjacent T2 comparisons (Table A18). For example, the T1 difference between InternVL3.5-38B and InternVL3.5-8B is 0.0015 with interval $[-0.0182, 0.0215]$. These comparisons do not resolve a definitive ordering; intervals containing zero do not establish model equivalence.

Larger contrasts remain separated: InternVL3.5-38B minus GPT-4o is 0.2473 $[0.2282, 0.2670]$ for T1; LISA-13B minus OWLv2-Large+SAM is 0.2135 $[0.1868, 0.2398]$ for T2. These pointwise intervals cover test-sampling uncertainty, excluding prompt, parsing, threshold, and API variation. Highlighting in the primary tables denotes point estimates; small differences require the uncertainty analysis above.

Table A18: Adjacent comparisons whose pointwise 95% bootstrap intervals contain zero. Each row reports the first model minus the second in T1 overall Set-F_1 at IoU 0.50 or T2 positive-query All-mIoU. Differences are computed before rounding and can differ by 0.0001 from subtraction of the displayed primary scores.

Task	First model − second model	Δ	Pointwise 95% CI
T1	InternVL3.5-38B − InternVL3.5-8B	0.0015	[−0.0182, 0.0215]
T1	Qwen3-VL-8B − DeepSeek-VL2-small	0.0080	[−0.0124, 0.0284]
T1	Florence-2-Large − Claude Sonnet 4.6	0.0024	[−0.0199, 0.0251]
T1	Claude Sonnet 4.6 − Gemini 2.5 Pro	0.0147	[−0.0088, 0.0373]
T2	LISA-13B − LISA-7B	0.0089	[−0.0082, 0.0260]
T2	LISA-7B − PixelLM-7B	0.0008	[−0.0191, 0.0206]
T2	PixelLM-7B − PSALM	0.0099	[−0.0130, 0.0327]
T2	PSALM − PixelLM-13B	0.0077	[−0.0138, 0.0293]
T2	GLaMM − Grounded-SAM	0.0177	[−0.0101, 0.0454]
T2	Florence-2-Large − SEEM	0.0007	[−0.0256, 0.0281]
T2	SAM 3 − OWLv2-Large+SAM	0.0120	[−0.0109, 0.0349]

Human-paraphrased queries. We manually paraphrase 100 image-disjoint T1 development queries, stratified by source and query type (50 single-target, 34 multi-target, and 16 target-absent). Semantic checks preserve category, cardinality, ranking or spatial relation, and target presence; for example, "the 3 leftmost trees" becomes "Which three trees are farthest to the left?" Only the query text changes within each pair, and no example is removed based on model outcomes. All four model–wording conditions yield 100 valid predictions without parsing or schema failures. Table A19 shows slightly higher positive-query Set-F_1 for both models under human wording, with their relative ordering preserved. GPT-4o returns fewer empty predictions on positive queries ($35/84$ to $25/84$), but correctly abstains on fewer target-absent queries ($13/16$ to $11/16$). These observations apply only to the two evaluated models and this T1 subset, rather than establishing general robustness to unrestricted language.

Table A19: Paired evaluation of original and human-paraphrased T1 queries. All, Pos., and Multi denote mean Set-F_1 at IoU 0.50 over 100 queries, 84 positive queries, and 34 multi-target queries, respectively. E-Acc uses 16 target-absent queries; Over-abst. is the fraction of the 84 positive queries answered with an empty set. All values are on a 0–1 scale.

Model	Wording	All	Pos.	Multi	E-Acc	Over-abst.
GPT-4o	Original	0.1500	0.0238	0.0000	0.8125	0.4167
GPT-4o	Human	0.1333	0.0278	0.0098	0.6875	0.2976
GPT-5.4	Original	0.2432	0.1585	0.1564	0.6875	0.0119
GPT-5.4	Human	0.2558	0.1736	0.1642	0.6875	0.0119

API run-to-run variation. We repeat GPT-4o and GPT-5.4 three times each on the same fixed, stratified sample of 400 T1 development queries, holding prompts, parsers, and decoding settings constant. Every run covers all 400 queries without missing or unparsable outputs. Table A20 reports individual scores and their mean and sample standard deviation. GPT-5.4 scores above GPT-4o in all three repetitions. This analysis measures API run-to-run variation on a fixed development sample, complementing the test-sampling uncertainty above; its scores are not full-test results or confidence intervals for other models.

Table A20: Repeated API evaluation on 400 fixed T1 development queries. Scores are overall Set-F_1 at IoU 0.50 on a 0–1 scale; SD is the sample standard deviation across three runs.

Model	Run 1	Run 2	Run 3	Mean ± SD
GPT-4o	0.1406	0.1450	0.1450	0.1435 ± 0.0025
GPT-5.4	0.2501	0.2632	0.2690	0.2608 ± 0.0097

E HOSTING, LICENSING, AND MAINTENANCE

Dataset hosting and access. AgroVG is publicly hosted on the Hugging Face Hub at `https://huggingface.co/datasets/sauryrs/AgroVG`, with the preprocessing, evaluation, and benchmarking code released at `https://anonymous.4open.science/r/AgroVG-5172/` during review and to be de-anonymized upon acceptance. The dataset release contains query files, dev/test splits, normalized metadata, mask provenance fields, evaluation inputs, released prediction files when permitted, and a Croissant metadata file describing dataset structure, field-level provenance, and split organization. An archival snapshot with a persistent identifier will be deposited upon acceptance.

Licensing. AgroVG follows the access and redistribution terms of each source dataset: every image and source-derived annotation inherits the license and usage conditions of its original dataset (Appendix B.1, Table A1). When raw-image redistribution is not permitted or is uncertain, AgroVG releases derived metadata, preprocessing scripts, annotation-grounded query records, and evaluation files, and users obtain raw images from the original source. AgroVG-specific contributions—the unified schema, derived masks, audit tables, query templates, split files, and evaluation protocols—are released under CC BY 4.0 where compatible with source terms, and the benchmarking code is released under the MIT license. A per-source license and attribution summary is included in the dataset card.

Maintenance plan. We will maintain AgroVG for at least three years after publication. Maintenance commitments include fixing reproducibility issues, releasing patched metadata when source-side errata are reported, documenting version history, and accepting community-contributed predictions for a public leaderboard. Issue tracking, contribution guidelines, and maintainer contacts are provided in the dataset card and code repository.