# GEOREFER-BENCH: A BENCHMARK FROM REFERRING PIXELS TO VERIFIABLE GEOSPATIAL REASONING

**Anonymous authors**
Paper under double-blind review

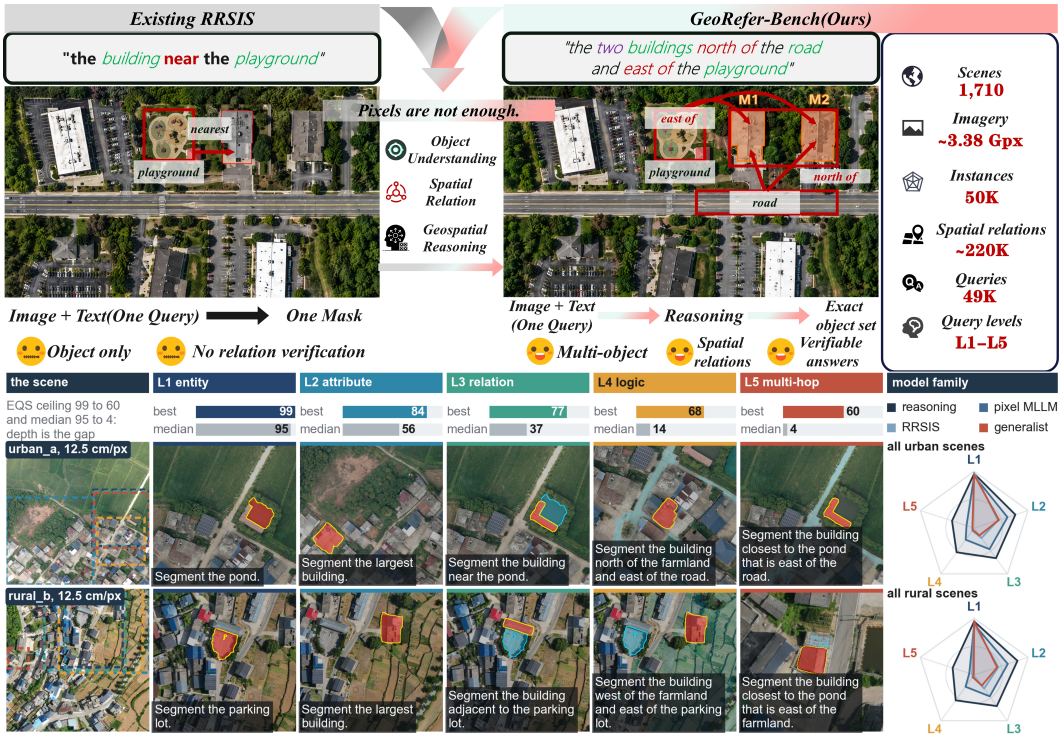


Figure 1: GeoRefer-Bench in one scene. *Top:* classical referring segmentation returns one mask for one salient object, while a sentence here names relations between objects and its answer is the set that satisfies them, scored by identity rather than by overlap; the panel on the right is the georeferenced UAV track in numbers. *Bottom:* one urban and one rural scene at five reasoning levels, each cell carrying the released sentence and its answer set, with the score band above and, on the right, the profile per method family.

## ABSTRACT

We introduce GeoRefer-Bench, a benchmark for evaluating whether vision-language models identify the objects denoted by spatial expressions in remote sensing imagery. Pixel overlap measures segmentation quality but does not directly assess whether the predicted object set satisfies a spatial query. GeoRefer-Bench links natural-language expressions to executable logical forms over annotated scene graphs, providing explicit query semantics and traceable answer sets. It comprises 1,710 scenes from georeferenced UAV imagery and three re-instrumented public datasets, with 50,488 objects and 49,512 executable queries covering five query structures, from entity selection to two-hop relations. Unanswerable queries, counterfactual pairs and paraphrases enable targeted tests of spatial reference resolution. We complement pixel overlap with Exact Query Success (EQS), assessing object selection through a shared mask-to-object read-out that uses ground-truth instance masks at evaluation time. Relation-blind baselines expose a discrepancy between overlap and correct selection: on answerable UAV relation queries, predicting all instances of the named category achieves 36.9 mIOU but only 9.2% EQS. Evaluation of fifteen models on 1,881 UAV test queries reveals substantial performance



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.

gaps between entity and compositional queries: the strongest overall model reaches 98.9 at entity selection but only 60.5 at two-hop composition. Stratified analyses also associate greater same-category ambiguity with lower success, even within answer-area strata. Together, these results motivate evaluating spatial referring segmentation through both pixel accuracy and query-consistent object selection.

## 1 INTRODUCTION

A spatial referring segmentation system must determine which objects a sentence denotes and then return their pixels. In overhead imagery, many objects share a category and similar appearance, so reference often depends on their relations to other objects or regions. For example, *the buildings north of the road* denotes a set whose size depends on the scene and may be zero. A useful evaluation must therefore examine both the identity of the selected objects and the quality of their boundaries. Strong performance on one requirement does not establish success on the other.

Referring segmentation has advanced through language–vision fusion (Hu et al., 2016; Ding et al., 2021; Yang et al., 2022; Wang et al., 2022) and mask-decoding multimodal language models (Lai et al., 2024; Rasheed et al., 2024). Remote-sensing datasets and models extend these capabilities to overhead imagery (Yuan et al., 2024; Liu et al., 2024; Kuckreja et al., 2024; Shabbir et al., 2025; Li et al., 2025b; Quenum et al., 2025), while generalized referring segmentation and NWPU-Refer already accommodate multiple targets and no-target expressions (Liu et al., 2023; Yang et al., 2025). Our focus is not the introduction of spatial language or variable answer cardinality. It is the connection between a sentence's spatial semantics, the objects that satisfy it, and the evidence used to evaluate a prediction.

Mask overlap remains useful for measuring delineation, but an aggregate overlap score does not state whether every required object, and only those objects, has been selected. Conversely, an exact selection score alone does not describe boundary quality. Resolving this distinction requires an object inventory and explicit query semantics, not merely another aggregate score. It also raises a second issue: when a model fails a spatial query, the failure may reflect visual resolution, candidate ambiguity, language processing or relation composition. A benchmark should expose these conditions without treating a low score as direct evidence for one internal failure mechanism.

**Two research questions. RQ1: How can spatial referring segmentation be benchmarked with explicit query semantics, traceable answer sets and verifiable evaluation?**

**RQ2: Under which query structures and scene conditions do current models fail to resolve spatial references?** To address these questions, GeoRefer-Bench combines executable query semantics and expert-reviewed scene annotations with a common object-set read-out. Relation-blind controls test the evaluation protocol; model comparisons diagnose query structure, candidate ambiguity, answerability and ranking stability.

**Contributions.**

1. **An executable protocol for spatial reference resolution.** We link sentences, spatial predicates and instance-level answer sets through executable logical forms, and evaluate exact object selection alongside pixel overlap. A shared read-out supports models that produce a single foreground mask, with its dependence on the annotated object inventory made explicit (Section 3; RQ1).
2. **A two-track benchmark with auditable construction.** We provide 700 georeferenced UAV scenes and 1,010 public remote-sensing scenes under the same grammar, totaling 50,488 objects, 219,990 relations and 49,512 executable queries (Table 1). Five query levels and three families of controls support structured evaluation. Domain experts repeatedly review reference resolution and annotations, including a human-traced boundary subset; programmatic audits check record consistency and provenance (Sections 4–5; RQ1).
3. **Diagnostics of selection errors and model limitations.** Relation-blind baselines on both tracks demonstrate why overlap must be interpreted separately from exact-set success. Evaluation of fifteen models on the UAV track identifies performance gaps across query

Table 1: From mask-level referring to verifiable geospatial reasoning. A *unit* is what one annotation is attached to. GeoRefer-Bench is one protocol on two tracks: georeferenced UAV orthophotos as whole 2048 scenes, and the published patches of RefSegRS, RRSIS-D and RISBench re-instrumented with the same grammar. *Progr.* counts executable logical forms and *sent.* their realizations; *objects / relations* the structure the release ships; *Metric* is a projected CRS, released on the public track only where a sampling is known (Appendix E).

| | released data | | | queries | | | | released structure | | | controls | | score |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| dataset | units | px/side | progr. | sent. | depth | multi | none | objects / relations | exec. | metric | para. | c-fact. | set |
| *Natural images* | | | | | | | | | | | | | |
| RefCOCO (2016) | 19,994 photos | var. | – | 142,209 | ○ | ○ | ○ | – | ○ | – | ◐ | ○ | ○ |
| gRefCOCO (2023) | 19,994 photos | var. | – | 278,232 | ○ | ● | ● | – | ○ | – | ○ | ○ | ○ |
| *Overhead imagery* | | | | | | | | | | | | | |
| RefSegRS (2024) | 4,420 patches | 512 | – | 4,420 | ○ | ○ | ○ | – | ○ | ○ | ○ | ○ | ○ |
| RRSIS-D (2024) | 17,402 patches | 800 | – | 17,402 | ◐ | ○ | ○ | – | ○ | ○ | ○ | ○ | ○ |
| RISBench (2024) | 52,472 patches | 512 | – | 52,472 | ◐ | ○ | ○ | – | ○ | ○ | ○ | ◐ | ○ |
| NWPU-Refer (2025) | 15,003 patches | ≤2048 | – | 49,745 | ◐ | ● | ● | – | ○ | ○ | ○ | ○ | ○ |
| *GeoRefer-Bench (ours)* | | | | | | | | | | | | | |
| UAV orthophotos | **700 scenes** | **2048** | **20,916** | **62,748** | ● | ● | ● | **26,217 / 142,796** | ● | ● | ● | ● | ● |
| Public RS referring | **1,010 scenes** | 512–800 | **28,596** | **85,788** | ● | ● | ● | **24,271 / 77,194** | ● | ◐ | ● | ● | ● |

● released as structure ◐ in the language only ○ absent – not applicable

structures and associations with same-category ambiguity, with scene-level uncertainty estimates and population reweighting controls (Sections 6–7; RQ1 and RQ2).

## 2 Related work

**Referring image segmentation.** Early referring segmentation benchmarks associate language expressions with target masks (Hu et al., 2016; Kazemzadeh et al., 2014; Yu et al., 2016; Mao et al., 2016). Subsequent methods improve language–vision interaction and exploit visual–language pretraining (Ding et al., 2021; Yang et al., 2022; Wang et al., 2022; Radford et al., 2021), including mask decoding from language models (Lai et al., 2024; Rasheed et al., 2024). Generalized referring segmentation permits multiple referents or no referent (Liu et al., 2023). These developments motivate our distinction between answer cardinality, object identity and pixel quality. We retain overlap as a segmentation diagnostic and add explicit object-set evaluation rather than treating the two measures as interchangeable.

**Referring and grounding in remote sensing.** Remote-sensing grounding includes box localization (Zhan et al., 2023), visual question answering (Lobry et al., 2020) and pixel-level reference resolution through RefSegRS (Yuan et al., 2024), RRSIS-D (Liu et al., 2024) and RISBench (Dong et al., 2024). NWPU-Refer extends coverage to higher-resolution imagery and multi- and no-target expressions (Yang et al., 2025). Multimodal remote-sensing models also support language-guided localization and segmentation (Hu et al., 2025; Kuckreja et al., 2024; Shabbir et al., 2025; Li et al., 2025b; Quenum et al., 2025). Our contribution is an executable evaluation layer: object inventories, spatial predicates and query programs make the intended selection inspectable. The public track re-instruments imagery from the first three corpora under this protocol; it does not replace their original annotations or report results on their original referring tasks (Appendix E).

**Program-based compositional evaluation.** CLEVR (Johnson et al., 2017) and GQA (Hudson & Manning, 2019), using annotated structure such as Visual Genome (Krishna et al., 2017), establish scene graphs and executable programs as useful tools for compositional evaluation. GeoRefer-Bench applies this principle to spatial reference resolution with mask-valued outputs and measured geometry. Detection, instance and land-cover resources provide complementary object or region annotations (Xia et al., 2018; Zamir et al., 2019; Wang et al., 2021), but a query additionally specifies which entities satisfy a relation. Promptable segmenters (Kirillov et al., 2023; Carion et al., 2026; Li et al., 2025a) support our annotation pipeline; the resolver selects answers from the resulting inventory. This makes answer derivation reproducible, but its correctness remains conditional on the inventory, geometry and predicate definitions.

## 3 TASK AND METRIC

To address RQ1, we specify the scene representation, executable query semantics and prediction evaluation separately. A scene is a tuple $S = (I, \mathcal{O}, \mathcal{R}, \mathcal{G})$. On the UAV track the image $I$ is a $2048 \times 2048$ orthorectified RGB tile; on the public track it is the published patch at its native size (512 or 800). $\mathcal{O}$ holds two kinds of entity: *discrete objects*, each with a pixel-disjoint binary mask $M_o$ and measured geometry (area in $\text{m}^2$ where a ground sampling is known, else in pixels; projected or image-plane centroid; bounding box), and *region anchors* such as road, water, forest or farmland, which may overlap and are stored in a bit-plane mask. $\mathcal{R}$ is a set of directed edges $(o_i, \rho, o_j)$. On the UAV track every edge is computed in a metric CRS: eight compass directions, the bearing quantized into $45^\circ$ sectors; two distance predicates, a metric bound (*within $d$ metres*) and proximity, in metres between projected geometries; and two topological predicates. On the public track the same families are used where metres exist; where they do not, $\text{WITHIN}(d)$ and the qualitative distance words are withheld rather than given an invented scale, and what remains — directions, topology, and proximity *ranking* — still reaches every reasoning level. $\mathcal{G}$ records the geodetic metadata a model may legitimately use: ground sampling distance when known, the CRS, and the direction of north.

A query is a pair $q = (T, \varphi)$: a sentence $T$ in English and a logical form $\varphi$ over $\mathcal{O}$ and $\mathcal{R}$ drawn from the grammar $\varphi ::= \text{ENT}(c) \mid \text{ATTR}(c, s) \mid \text{REL}(c, \rho, a) \mid \text{AND}(c, \rho_1, a_1, \rho_2, a_2) \mid \text{HOP}(c, \rho_1, a_1, \rho_2, a_2)$, with $c$ a target category, $s$ a superlative, $\rho$ a spatial predicate and $a$ an anchor. ATTR holds only the measured extremes that need no second object — largest, smallest, northernmost, southernmost — so a proximity selector, which always names one, is REL or HOP rather than ATTR; and the five productions are the five reasoning levels of Table 2. Every predicate has a deterministic implementation. Directions, topology, the metric bound $\text{WITHIN}(d)$ (46–92 m in the sentence) and proximity ranking are defined on the stored geometry; their physical definitions do not depend on pixel sampling, although the estimated geometry can change with sampling or annotation error. Only *near* and *far* are calibrated to the ground window (5.2% of UAV queries), flagged so a cross-scale reading can exclude them when checking sensitivity (Appendix M). The answer set is defined by execution,

$$A(q) \;=\; \text{RESOLVE}(\varphi, \mathcal{O}, \mathcal{R}) \;\subseteq\; \mathcal{O}, \qquad |A(q)| \;\geq 0, \tag{1}$$

and $A(q) = \emptyset$ is a legal answer. The generator, validator and auditor use the same RESOLVE implementation, so stored answers can be checked against the released logical forms and scene records. This establishes internal execution consistency. Complementary domain-expert review examines whether the language and visual evidence support the recorded referents (Section 5); execution checks alone cannot detect every shared implementation error, missed object or linguistic ambiguity. A model receives $(I, T, \mathcal{G})$ where its interface supports the metadata, never $\varphi$ or $\mathcal{R}$, and returns binary masks $P = \{P_1, \ldots, P_m\}$, with $m = 0$ permitted.

**Exact Query Success.** Let $\tau = 0.5$ and let $A(q) = \{a_1, \ldots, a_k\}$ with masks $M_{a_1}, \ldots, M_{a_k}$. Then

$$\text{EQS}(P, q) = \begin{cases} \mathbb{1}[\, m = 0 \,] & \text{if } k = 0, \\ \mathbb{1}\big[\, m = k \;\wedge\; \exists\, \pi \in \mathfrak{S}_k :\; \text{IoU}(P_i, M_{\pi(i)}) \geq \tau \;\; \forall i \,\big] & \text{if } k > 0, \end{cases} \tag{2}$$

where the matching $\pi$ is obtained by Hungarian assignment on the IoU matrix. For native instance outputs, Equation 2 requires the correct number of instances and a qualifying match for each answer object. Thus it jointly tests selection and delineation at threshold $\tau$. A single union mask cannot satisfy a multi-instance answer under this native-output definition; the read-out below provides a separate selection-level evaluation for such outputs.

**Reading a single mask as a set.** Most current systems emit one mask per sentence, which would score zero on every multi-object answer for a reason unrelated to reasoning. The read-out is therefore part of the protocol: writing $P^{\cup} = \bigcup_i P_i$,

$$\widehat{A}(P) = \big\{\, o \in \mathcal{O} \;:\; |P^{\cup} \cap M_o| \,/\, |M_o| \;\geq\; \eta \,\big\}, \qquad \eta = 0.5, \tag{3}$$

which identifies the annotated objects sufficiently covered by the predicted foreground. Under this read-out, exact selection on an answerable query means that $\widehat{A}(P)$ equals $A(q)$; count accuracy and set-based instance F1 use the same identities. All systems in Section 7 use this common read-out, including those with multiple output masks. The reported EQS therefore measures object selection

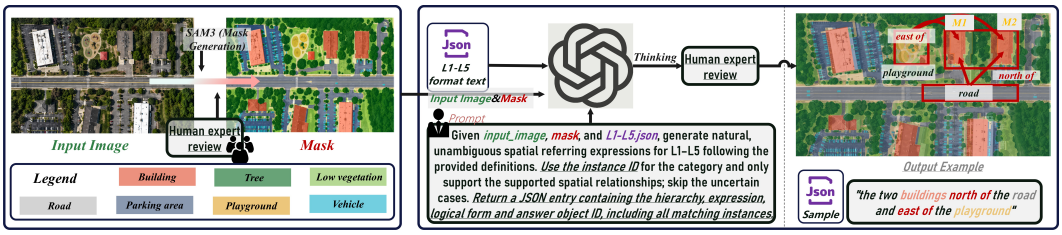


Figure 2: Benchmark construction. SAM3 proposes masks for expert inspection and correction. The image, reviewed masks and L1–L5 specification guide LLM-assisted formulation of spatial expressions and candidate logical forms, which experts verify against the scene. Accepted forms are executed on the annotated scene graph to obtain answer sets and query-specific masks; non-target instances become background only in the query mask. The illustration shows the workflow, not the complete category vocabulary of either track.

relative to the ground-truth inventory, rather than native instance separation. It is not equivalent to applying the IoU matching rule in Equation 2 to raw predictions: foreground outside annotated objects may not affect the read-out. Pixel overlap must consequently be reported alongside it. For unanswerable queries, the protocol instead checks the raw output; any non-empty foreground is a rejection failure.

**Secondary metrics.** The scorer additionally provides mIoU over answerable queries, instance F1 and count accuracy, $|\widehat{A}(P)| = |A(q)|$, together with per-predicate accuracy, counterfactual consistency (both members of a pair exactly correct), rejection accuracy on unanswerable queries, and paraphrase consistency (all three wordings of a form exactly correct). Native instance outputs can be matched as in Equation 2; the common mask read-out uses the selected object identities instead. We report EQS over all queries and over answerable queries separately. This matters because always returning an empty prediction succeeds on $24.0\%$ of the full UAV query collection; that proportion is specific to this collection and need not equal the evaluation sample's proportion.

**What one query is.** The unit is the pair $(\varphi, A(q))$, not the sentence: the UAV track holds $20{,}916$ logical queries and the public track $28{,}596$, each realized in three semantically equivalent wordings, hence $62{,}748$ and $85{,}788$ query–sentence pairs. Of those, $14{,}797$ and $28{,}827$ are distinct strings because controlled linguistic realizations recur across scenes. Systems in Section 7 are scored on one canonical realization per logical query of the UAV track, so the $1{,}881$ test items of Table 3 are $1{,}881$ logical queries and the other realizations enter only through paraphrase consistency.

## 4 Constructing the Benchmark

Figure 2 summarizes construction: mask proposals, expert verification, LLM-assisted query formulation and executable answer derivation. Both imagery tracks use the scene representation and grammar of Section 3; their coverage is shown in Figure 5 in Appendix D.

**Imagery, scenes and instances.** Seven Jiangxi orthophotos supply 700 UAV scenes at two ground samplings (Appendix D). SAM3 and text-promptable segmentation (Carion et al., 2026; Li et al., 2025a) provide initial proposals. Boundary evidence, LLM-assisted checks and repeated domain-expert review verify the annotations; a human-traced subset supports boundary inspection (Appendix G). The two samplings differ in physical field of view and are not a matched resolution experiment.

**Public referring corpora.** The public track applies the same protocol to $1{,}010$ RefSegRS, RRSIS-D and RISBench scenes, inheriting their official splits. Original expressions are retained as provenance; evaluation uses our executable queries. RefSegRS metric quantities use an estimated scale, while the other subsets retain image-plane geometry (Appendix E).

**Geometry-grounded scene graphs.** Spatial predicates are computed from reviewed geometry under the coordinate conventions of Section 3; LLM-proposed relations are checked against this

representation. To make anchors identifiable, a discrete reference must be the sole instance of its category, and a region covering more than $55\%$ of the scene cannot serve as an anchor.

**Queries, controls and paraphrases.** Following program-based construction (Johnson et al., 2017; Hudson & Manning, 2019), the L1–L5 specification constrains LLM-assisted proposals of expressions and logical forms. Experts review their spatial validity and reference clarity; retained forms define selection semantics, while controlled language realization preserves their correspondence to the expressions. The five structures in Table 2 describe program composition, not equal increments of cognitive difficulty. Both tracks include absent-category or unsatisfied-relation queries, counterfactual pairs that change one relation and its answer, and three equivalent wordings per form. Equation 1 determines the final answer sets rather than accepting LLM-proposed object IDs as ground truth. Query masks retain the selected instances and assign other pixels to background; experts review expression–answer–mask consistency before release.

Table 2 summarizes the query composition and relation-blind diagnostics.

**Difficulty and splits.** Difficulty tiers combine candidate count, query structure, answer area and answer cardinality (Appendix I). The primary UAV split holds out the southernmost $20\%$ of each orthophoto and removes overlapping northern scenes to prevent native-pixel leakage. The public track inherits official assignments; additional geographic, scale and compositional protocols are detailed in Appendix B.

## 5 Validation and Scope of the Audit

Validation combines expert review with computational checks. Domain experts repeatedly inspect spatial reference resolution and visual annotations, with a human-traced subset for boundary review (Appendix G). Public-source masks provide an additional comparison at their original annotation granularity (Appendix E).

The programmatic audit re-executes queries, checks mask–record consistency, reconstructs native-window provenance and measures tile-edge artifacts. An initial build yielded 86 issues over 10 scenes; the revised rules yield 0 detected issues across the released 1,710 scenes, with provenance recovered for all 700 UAV windows (Appendix J). This count describes compliance with automated checks, not human agreement or exhaustive annotation correctness. Residual mask or inventory errors can still propagate into spatial answers.

Table 2: What each reasoning level is, and what it currently costs. *Cand.* is the mean number of same-category instances a sentence must choose among, *area%* the median share of the scene its answer covers. The right-hand block is measured (Section 6): three relation-blind strategies handed the ground-truth masks, so only selection is tested. Its level rows are EQS over the answerable queries of that level, its total row EQS over all 20,916 UAV-track queries, which is why a strategy that always returns something scores below its own level rows; the last column is mIOU for the category strategy — 36.9 at L3 against an exact-set 9.2.

| reasoning level | | | queries | | | answers and language | | | relation-blind EQS | | | overlap |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | selects | logical form | total | unans. | ans./q | cand. | area% | words | categ. | largest | random | categ. mIOU |
| **L1** | entity | ENT($c$) | 2,384 | 1,400 | 2.97 | 3.0 | 1.12 | 3.1 | 100.0 | 71.5 | 71.5 | 100.0 |
| the category alone; every instance of it answers "Segment the pond." | | | | | | | | | | | | |
| **L2** | superlative | ATTR($c$, $sup$) | 2,760 | 0 | 1.00 | 34.7 | 0.12 | 4.0 | 0.0 | 30.4 | 10.1 | 12.2 |
| a measured extreme, no anchor: a size prior nearly works "Segment the largest building." | | | | | | | | | | | | |
| **L3** | relation | REL($c$, $\rho$, $a$) | 10,473 | 2,714 | 2.67 | 22.7 | 0.52 | 7.5 | 9.2 | 16.9 | 15.6 | 36.9 |
| one relation to a named anchor, and the category stops sufficing "Segment the ship near the road." | | | | | | | | | | | | |
| **L4** | conjunction | AND($\rho_1$, $\rho_2$) | 3,646 | 907 | 1.99 | 47.0 | 0.26 | 12.4 | 0.0 | 5.5 | 3.6 | 11.8 |
| two relations, both true of the same instance "Segment the greenhouses south of the parking lot and south of the road." | | | | | | | | | | | | |
| **L5** | two-hop | HOP($c$, $a_1 \to a_2$) | 1,653 | 0 | 1.00 | 37.0 | 0.16 | 13.2 | 0.0 | 13.4 | 13.6 | 13.3 |
| the anchor is itself picked by a relation first "Segment the building closest to the pond that is east of the road." | | | | | | | | | | | | |
| **all queries** | | | 20,916 | 5,021 | 2.11 | 29.2 | 0.33 | 7.8 | 16.0 | 22.7 | 19.0 | 29.7 |

Table 3: Fifteen models on the UAV track of GeoRefer-Bench, over the 1,881 test queries, in per cent: EQS over all queries and over answerable ones, mIoU, rejection accuracy on the unanswerable sentences, and EQS per reasoning level and per difficulty tier, because a single mean hides which level failed. $\Delta_{1\to5}$ is the L1-to-L5 drop, the only column where small is good. A red ↓ on the answerable column marks a system whose headline number is held up by the sentences whose right answer is nothing. Best per column in **red**, runner-up underlined; references and inference settings in Appendix L.

| method | venue | EQS (%) | | mIoU | rej. | EQS per level, answerable | | | | | $\Delta_{1\to5}$ | EQS per tier | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | all | ans. | | | L1 | L2 | L3 | L4 | L5 | | easy | med. | hard |
| *Remote-sensing reasoning segmentation* | | | | | | | | | | | | | | |
| SegEarth-R2 | CVPR'26 | **74.1** | **77.0** ↑ | **71.2** | 62.8 | 98.9 | **84.3** | **77.5** | **67.8** | **60.5** | **38.4** | 91.7 | **77.3** | **71.0** |
| Geo-R1 | JPRS'26 | 73.2 | 75.2 ↑ | 66.1 | **65.4** | 98.3 | 84.0 | 74.9 | 66.1 | 56.5 | 41.9 | **93.1** | 74.5 | 68.5 |
| UniGeoSeg | CVPR'26 | 70.1 | 72.8 ↑ | 64.9 | 60.0 | **99.4** | 83.2 | 74.1 | 61.7 | 48.8 | 50.6 | 92.0 | 73.5 | 64.7 |
| LISAt | NeurIPS'25 | 59.6 | 60.1 ↑ | 45.6 | 57.7 | 92.7 | 75.6 | 58.3 | 45.3 | 33.5 | 59.2 | 82.3 | 58.9 | 52.1 |
| SegEarth-R1 | arXiv'25 | 44.7 | 44.2 ↓ | 29.9 | 46.7 | 88.2 | 60.4 | 46.9 | 26.5 | 4.8 | 83.4 | 77.1 | 43.4 | 31.7 |
| *Remote-sensing foundation and grounding models* | | | | | | | | | | | | | | |
| GeoPixel | ICML'25 | 59.5 | 60.5 ↑ | 43.3 | 55.9 | 96.6 | 72.7 | 59.3 | 48.0 | 32.7 | 64.0 | 87.8 | 61.5 | 49.0 |
| RemoteSAM | ACM MM'25 | 39.9 | 39.2 ↓ | 30.1 | 42.6 | 91.0 | 60.6 | 37.3 | 13.8 | 2.4 | 88.6 | 77.1 | 34.9 | 26.9 |
| GeoPix | GRSM'25 | 39.9 | 38.2 ↓ | 24.9 | 46.2 | 86.0 | 52.8 | 42.2 | 15.1 | 3.2 | 82.7 | 73.3 | 36.0 | 25.8 |
| *Classical RRSIS* | | | | | | | | | | | | | | |
| RS2-SAM2 | AAAI'26 | 38.3 | 37.7 ↓ | 33.2 | 40.8 | 94.9 | 55.9 | 37.3 | 8.1 | 4.8 | 90.1 | 75.7 | 34.9 | 24.5 |
| RSRefSeg2 | TGRS'26 | 37.1 | 36.7 ↓ | 30.6 | 38.7 | 97.2 | 56.4 | 36.3 | 3.7 | 3.2 | 94.0 | 79.9 | 33.1 | 22.0 |
| FIANet | TGRS'25 | 29.7 | 28.4 ↓ | 26.0 | 34.9 | 95.5 | 43.3 | 17.6 | 4.4 | 2.8 | 92.7 | 72.2 | 23.2 | 14.4 |
| CADFormer | JSTARS'25 | 27.7 | 26.2 ↓ | 23.3 | 33.6 | 95.5 | 38.3 | 13.5 | 4.7 | 3.2 | 92.3 | 68.4 | 22.3 | 12.0 |
| RMSIN | CVPR'24 | 25.5 | 23.6 ↓ | 19.7 | 32.8 | 92.7 | 28.3 | 15.0 | 3.7 | 4.0 | 88.7 | 68.4 | 18.1 | 9.5 |
| *Generalist segmentation MLLM* | | | | | | | | | | | | | | |
| X2SAM | ECCV'26 | 32.4 | 30.2 ↓ | 23.8 | 40.8 | 93.8 | 48.0 | 20.7 | 5.0 | 2.4 | 91.4 | 72.6 | 25.7 | 16.5 |
| X-SAM | AAAI'26 | 29.9 | 27.2 ↓ | 21.4 | 40.3 | 94.4 | 42.5 | 15.0 | 3.7 | 2.4 | 92.0 | 72.2 | 21.3 | 13.2 |

Table 4: Fifteen models on the public track of GeoRefer-Bench, in per cent: the official test assignments of the three re-instrumented corpora (40 RefSegRS, 90 RRSIS-D and 142 RISBench scenes, 7,879 canonical queries). Per source, $EQS_a$ is exact query success over all its official-test queries and L1–L5 are answerable-query EQS at that level. Best per column in **red**, runner-up underlined; row order and venues as in Table 3. Answerable queries per level — RefSegRS 50/160/493/176/47, RRSIS-D 218/348/824/309/105, RISBench 385/536/1,343/575/267 — leave the RefSegRS L1 and L5 columns the thinnest. Per-source mIoU and answerable EQS, and the population actually scored, are Appendix F.

| method | RefSegRS (2024) | | | | | | RRSIS-D (2024) | | | | | | RISBench (2024) | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | $EQS_a$ | L1 | L2 | L3 | L4 | L5 | $EQS_a$ | L1 | L2 | L3 | L4 | L5 | $EQS_a$ | L1 | L2 | L3 | L4 | L5 |
| *Remote-sensing reasoning segmentation* | | | | | | | | | | | | | | | | | | |
| SegEarth-R2 | 73.1 | 98.0 | **88.1** | 76.3 | 61.4 | **55.3** | **74.6** | 98.6 | 82.5 | **78.4** | 64.1 | **56.2** | 72.3 | 98.2 | **87.7** | **77.8** | 61.4 | 50.2 |
| Geo-R1 | **74.7** | **100.0** | 85.6 | **78.7** | **67.0** | 53.2 | 73.6 | 98.6 | **85.3** | 75.1 | **67.0** | 51.4 | **73.2** | 98.7 | 84.0 | 77.4 | **68.2** | **58.1** |
| UniGeoSeg | 72.2 | 98.0 | 85.6 | 75.7 | 59.7 | 51.1 | 71.4 | **99.1** | 82.2 | 75.8 | 60.2 | 42.9 | 69.5 | **99.0** | 79.1 | 73.6 | 60.5 | 51.3 |
| LISAt | 59.9 | 98.0 | 73.1 | 63.3 | 49.4 | 27.7 | 62.6 | 95.9 | 76.4 | 64.9 | 45.6 | 33.3 | 60.5 | 96.9 | 75.4 | 64.6 | 46.1 | 30.3 |
| SegEarth-R1 | 45.3 | **100.0** | 56.9 | 46.7 | 21.6 | 12.8 | 47.7 | 93.1 | 63.2 | 48.5 | 20.4 | 6.7 | 49.8 | 94.0 | 66.6 | 53.9 | 27.0 | 5.6 |
| *Remote-sensing foundation and grounding models* | | | | | | | | | | | | | | | | | | |
| GeoPixel | 59.6 | 96.0 | 71.9 | 63.5 | 49.4 | 34.0 | 62.8 | 94.5 | 75.9 | 66.6 | 50.8 | 28.6 | 61.5 | 94.8 | 77.8 | 65.3 | 43.8 | 39.0 |
| RemoteSAM | 39.9 | 94.0 | 66.2 | 39.4 | 11.9 | 6.4 | 44.0 | 95.0 | 60.9 | 44.7 | 12.9 | 2.9 | 42.6 | 97.4 | 57.5 | 46.1 | 13.2 | 3.0 |
| GeoPix | 40.2 | 86.0 | 62.5 | 38.7 | 14.2 | 2.1 | 44.8 | 94.5 | 60.1 | 46.0 | 16.5 | 1.9 | 43.9 | 91.2 | 59.0 | 47.1 | 19.7 | 2.2 |
| *Classical RRSIS* | | | | | | | | | | | | | | | | | | |
| RS2-SAM2 | 37.0 | 96.0 | 57.5 | 35.3 | 8.5 | 2.1 | 42.3 | 96.3 | 55.7 | 44.2 | 9.4 | 1.9 | 40.1 | 96.4 | 57.6 | 42.9 | 8.0 | 2.6 |
| RSRefSeg2 | 33.5 | 94.0 | 49.4 | 33.3 | 4.5 | 4.3 | 39.6 | 95.0 | 58.9 | 39.6 | 5.2 | 1.0 | 37.9 | 97.9 | 51.5 | 40.1 | 4.9 | 2.6 |
| FIANet | 24.7 | 98.0 | 38.1 | 18.9 | 3.4 | 2.1 | 31.0 | 97.2 | 45.7 | 22.7 | 3.9 | 1.9 | 30.6 | 96.1 | 43.7 | 26.5 | 3.1 | 3.0 |
| CADFormer | 23.2 | 96.0 | 42.5 | 14.0 | 5.1 | 0.0 | 29.5 | 97.2 | 42.2 | 19.9 | 3.2 | 3.8 | 28.8 | **99.0** | 41.2 | 20.8 | 4.7 | 3.0 |
| RMSIN | 23.2 | 94.0 | 38.1 | 14.2 | 6.8 | 6.4 | 26.7 | 94.5 | 31.0 | 17.7 | 3.9 | 1.9 | 27.2 | 95.1 | 33.2 | 19.7 | 4.7 | 3.0 |
| *Generalist segmentation MLLM* | | | | | | | | | | | | | | | | | | |
| X2SAM | 28.4 | 96.0 | 47.5 | 20.7 | 4.5 | 2.1 | 33.4 | **99.1** | 42.8 | 24.3 | 3.9 | 2.9 | 33.2 | 97.7 | 45.5 | 27.3 | 3.7 | 0.7 |
| X-SAM | 26.0 | 96.0 | 38.8 | 19.5 | 4.0 | 0.0 | 31.5 | 96.3 | 39.1 | 23.9 | 3.2 | 1.9 | 31.6 | 96.4 | 45.5 | 24.9 | 4.0 | 2.2 |

## 6 Testing the evaluation protocol with relation-blind controls

To test whether overlap distinguishes spatial selection from category recognition, three privileged controls receive ground-truth masks and select all instances of the named category, the largest instance, or a random instance at a fixed seed. These controls remove segmentation error and test selection alone.

**Selection and overlap measure different outcomes.** On answerable UAV L3 queries, all-category selection achieves 36.9 mIoU but only 9.2 EQS (Table 2). It succeeds on every answerable L1 query and none at L2, L4 or L5; across all UAV queries, no control exceeds 22.7 EQS. Thus overlap

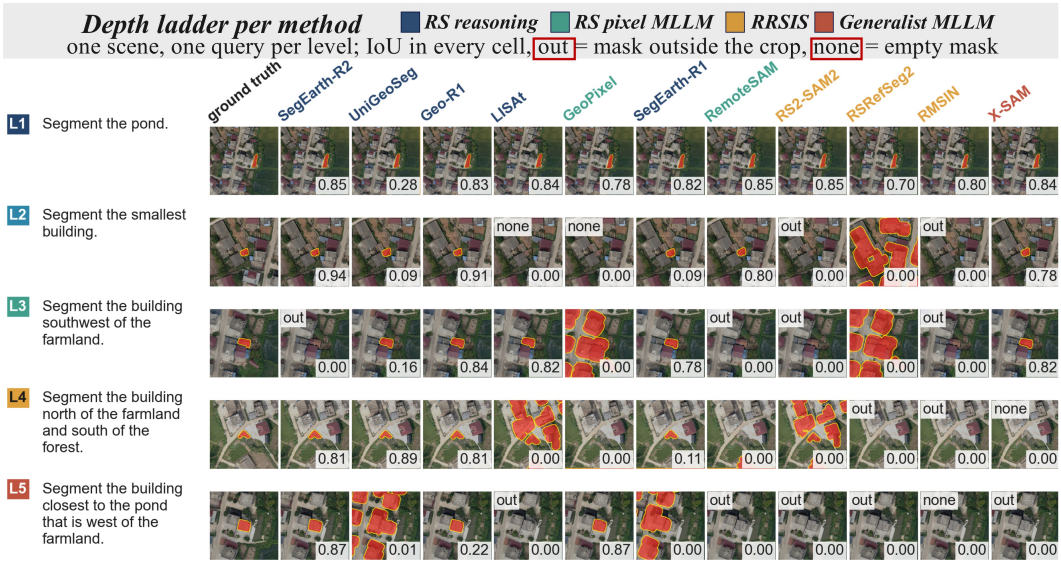


Figure 3: The depth ladder: one scene, one released query per reasoning level, eleven of the fifteen evaluated models in family colour (four dropped for legibility at page width). Cells carry IoU; *out* is a mask outside the crop and *none* an empty prediction. Read downwards: a column holds while the sentence names a category or a superlative and comes apart once it names a relation.

can coexist with incorrect selection. The metrics describe complementary outcomes, with mIoU restricted to answerable queries and overall EQS including empty answers.

**The diagnostic extends to public imagery.** All-category selection on public imagery similarly achieves 37.4 mIoU and 20.2 all-query EQS, with 10.0 answerable-query EQS at L3 (Appendix E). The discrepancy therefore extends beyond UAV imagery under the shared grammar, motivating exact-set evaluation alongside overlap.

## 7 Diagnosing Current Models

**Evaluation scope.** We run inference for all fifteen comparison models from four families using their respective released or officially provided checkpoints on 1,881 queries from 130 UAV test scenes. Each receives the image and sentence through its official interface, with metadata where supported, but never the logical form. All outputs use Equation 3. Table 3 reports the UAV results; Table 4 the same systems on the official test queries of the three public sources, whose protocol is detailed in Appendix F. Configurations are detailed in Appendix L.

**Query structure: category naming does not imply relational success.** All models reach at least 86.0 EQS on answerable L1 queries, but ten score below 5 at L5. SegEarth-R2, the strongest overall system, achieves 74.1 overall and declines from 98.9 at L1 to 60.5 at L5 (Table 3). The same decline repeats on all three public sources (Table 4), so the depth profile is not an artefact of the UAV orthophotos. Figures 3 and 4 illustrate these selection failures. Entity-level success therefore does not extend reliably to composed references; the comparison concerns configured systems, not isolated reasoning components.

**Candidate ambiguity and answer area.** Within L3, mean success across models falls from 49.0 with at most five same-category candidates to 34.1 with more than fifty. Answer-area stratification retains a 9.7–17.4-point gap (Appendix M), supporting ambiguity as a diagnostic beyond marginal answer area. These are associations: category and scene differences remain, and sentence length covaries with composition.

**Answerability and wording.** For ten models, all-query EQS exceeds answerable-query EQS, reflecting stronger performance on empty answers. Both aggregates and rejection accuracy are

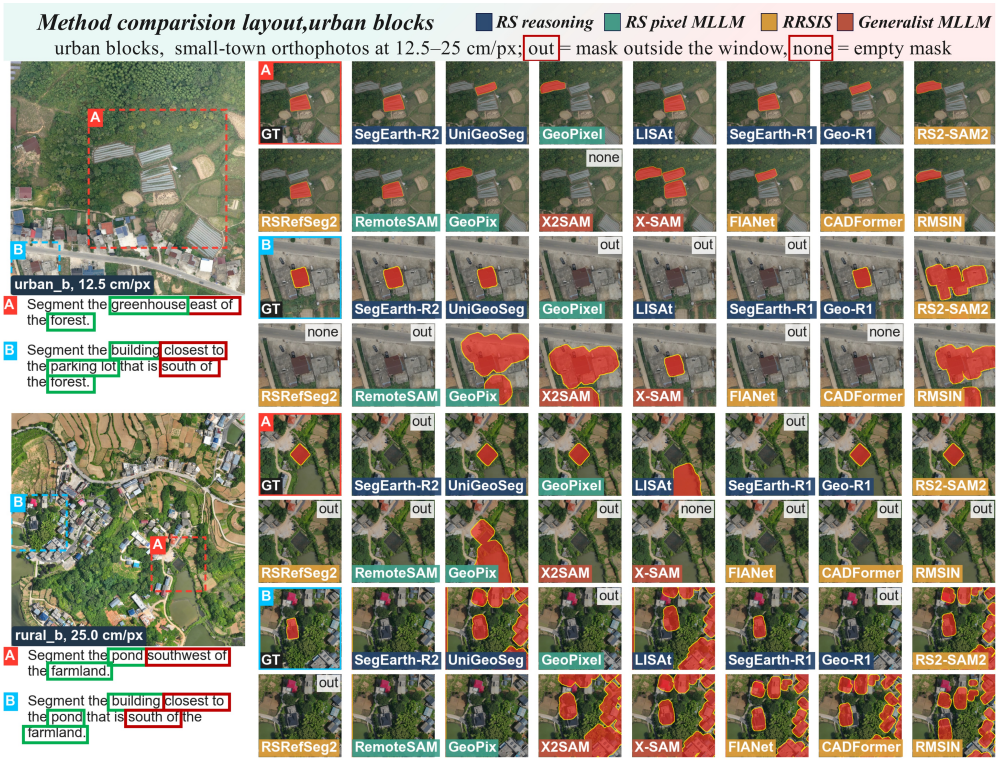


Figure 4: Predictions across scenes and ground samplings, one fine and one coarse scene: the whole scene with two dashed windows and its sentence, then those windows across the ground truth and all fifteen evaluated models. *Out* marks a mask outside the window, *none* an empty prediction — the two ways a wrong answer hides in a crop. Further scenes: Figure 8.

therefore reported. Error and paraphrase galleries expose selection failures and wording sensitivity without estimating their population prevalence (Appendices N and O).

**Reliability of the comparisons.** A paired scene bootstrap with 10,000 draws gives overall EQS intervals of approximately $\pm 2$ points. Multiple-comparison correction leaves the top-three ordering unresolved. Reweighting supported categories and excluding window-calibrated queries yield Kendall's $\tau = 0.96$ and 0.99, respectively (Appendix M). Broad rankings are stable under these checks, while small score differences warrant caution.

## 8 Discussion and Conclusion

**What the two questions establish.** RQ1 yields an inspectable link from spatial language to object sets, supported by expert review and execution checks. RQ2 reveals weaknesses in composed references and crowded candidate sets. Together, these findings motivate evaluating query-consistent selection alongside delineation.

**Limitations.** Public-source evaluation broadens imagery coverage under the shared grammar, but does not establish unrestricted-language or UAV-to-public transfer performance. Expert review reduces, but cannot eliminate, annotation, interpretation and scale-estimation errors.

**Conclusion.** GeoRefer-Bench makes spatial referring segmentation evaluable as both object-set selection and pixel delineation. Its executable queries and diagnostic controls expose selection errors that overlap alone does not characterize. Across UAV and public imagery, the results identify relational composition as a persistent challenge; within-level analyses further associate candidate ambiguity with lower success. The benchmark supports progress that can be assessed separately in selecting the right objects and delineating their boundaries.

### Reproducibility statement

The release includes scene images, instance and region masks, scene graphs, executable queries, paraphrases, diagnostic records and split assignments, together with the construction and evaluation scripts. Dataset statistics and selection controls correspond to Table 1, Table 2 and Appendices K and E. Model comparisons, uncertainty estimates and visualizations are derived from stored per-query predictions, with checkpoint and inference configurations described in Appendix L and the accompanying configuration files. Public-source model tables use the official test assignment (Table 4; Appendix F). Appendix M details ranking analyses. The validator checks record consistency, and UAV split provenance is reconstructed from native image windows. These resources support reproduction of the reported protocol. The expert-review procedure and human-traced boundary subset are described in Appendix G; automated consistency checks and human validation address complementary aspects of annotation quality.

### Ethics statement

The UAV-track imagery is UAV orthophotography of rural and urban land in Jiangxi, China, collected for land-survey purposes, and it is released for research use in scenes of $2048^2$ pixels at $0.125$–$0.25$ m ground sampling. At those ground sampling distances faces are not resolved and we know of no way to identify an individual from the released pixels, though we claim no formal anonymity guarantee. That track's vocabulary contains no person or vehicle category — vehicles were excluded because their segmentation was not stable enough across scenes and resolutions, which also removes the most identifying content the imagery could carry. The public track re-releases imagery already published by RefSegRS, RRSIS-D and RISBench, under those corpora's terms, and it does include vehicles at the coarser sampling of the sources, where faces remain unresolved. Sentences are generated from geometry, so they contain no annotator-supplied judgements about the places depicted. We note the dual-use character of any building-level geospatial resource: the release describes structures, not occupants, and it carries no attributes beyond category and geometry.

### AI use statement

LLMs assisted annotation checks and candidate expression and logical-form construction, as described in Section 4. Domain experts reviewed these outputs, and executable queries determined the released answer sets.

## References

Nicolas Carion, Laura Gustafson, Yuan-Ting Hu, Shoubhik Debnath, Ronghang Hu, Didac Suris Coll-Vinent, Chaitanya Ryali, Kalyan Vasudev Alwala, Haitham Khedr, Andrew Huang, Jie Lei, Tengyu Ma, Baishan Guo, Arpit Kalla, Markus Marks, Joseph Greer, Meng Wang, Peize Sun, Roman Rädle, Triantafyllos Afouras, 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 Cheng, Piotr Dollár, Nikhila Ravi, Kate Saenko, Pengchuan Zhang, and Christoph Feichtenhofer. SAM 3: Segment anything with concepts. In *International Conference on Learning Representations (ICLR)*, 2026.

Keyan Chen, Chenyang Liu, Bowen Chen, Jiafan Zhang, Zhengxia Zou, and Zhenwei Shi. RSRefSeg 2: Decoupling referring remote sensing image segmentation with foundation models. *IEEE Transactions on Geoscience and Remote Sensing*, 64:1–20, 2026. doi: 10.1109/TGRS.2025.3647535.

Henghui Ding, Chang Liu, Suchen Wang, and Xudong Jiang. Vision-language transformer and query generation for referring segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 16321–16330, 2021.

Zhe Dong, Yuzhe Sun, Tianzhu Liu, Wangmeng Zuo, and Yanfeng Gu. Cross-modal bidirectional interaction model for referring remote sensing image segmentation. arXiv:2410.08613 [cs.CV], 2024.

Ronghang Hu, Marcus Rohrbach, and Trevor Darrell. Segmentation from natural language expressions. In *European Conference on Computer Vision (ECCV)*, pp. 108–124, 2016.

Yuan Hu, Jianlong Yuan, Congcong Wen, Xiaonan Lu, Yu Liu, and Xiang Li. RSGPT: A remote sensing vision language model and benchmark. *ISPRS Journal of Photogrammetry and Remote Sensing*, 224:272–286, 2025. doi: 10.1016/j.isprsjprs.2025.03.028.

Drew A. Hudson and Christopher D. Manning. GQA: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6700–6709, 2019.

Justin Johnson, Bharath Hariharan, Laurens van der Maaten, Li Fei-Fei, C. Lawrence Zitnick, and Ross Girshick. CLEVR: A diagnostic dataset for compositional language and elementary visual reasoning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2901–2910, 2017.

Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara Berg. ReferItGame: Referring to objects in photographs of natural scenes. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 787–798, 2014.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 4015–4026, 2023.

Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A. Shamma, Michael S. Bernstein, and Li Fei-Fei. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International Journal of Computer Vision*, 123(1):32–73, 2017.

Kartik Kuckreja, Muhammad Sohail Danish, Muzammal Naseer, Abhijit Das, Salman Khan, and Fahad Shahbaz Khan. GeoChat: Grounded large vision-language model for remote sensing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 27831–27840, 2024.

Xin Lai, Zhuotao Tian, Yukang Chen, Yanwei Li, Yuhui Yuan, Shu Liu, and Jiaya Jia. LISA: Reasoning segmentation via large language model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9579–9589, 2024.

Sen Lei, Xinyu Xiao, Tianlin Zhang, Heng-Chao Li, Zhenwei Shi, and Qing Zhu. Exploring fine-grained image-text alignment for referring remote sensing image segmentation. *IEEE Transactions on Geoscience and Remote Sensing*, 63:1–11, 2025. doi: 10.1109/TGRS.2024.3522293.

Kaiyu Li, Ruixun Liu, Xiangyong Cao, Xueru Bai, Feng Zhou, Deyu Meng, and Zhi Wang. SegEarth-OV: Towards training-free open-vocabulary segmentation for remote sensing images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10545–10556, 2025a.

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. arXiv:2504.09644 [cs.CV], 2025b.

Chang Liu, Henghui Ding, and Xudong Jiang. GRES: Generalized referring expression segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 23592–23601, 2023.

Maofu Liu, Xin Jiang, and Xiaokang Zhang. CADFormer: Fine-grained cross-modal alignment and decoding transformer for referring remote sensing image segmentation. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 18:14557–14569, 2025. doi: 10.1109/JSTARS.2025.3576595.

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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 26658–26668, 2024.

Sylvain Lobry, Diego Marcos, Jesse Murray, and Devis Tuia. RSVQA: Visual question answering for remote sensing data. *IEEE Transactions on Geoscience and Remote Sensing*, 58(12):8555–8566, 2020. doi: 10.1109/TGRS.2020.2988782.

Junhua Mao, Jonathan Huang, Alexander Toshev, Oana Camburu, Alan L. Yuille, and Kevin Murphy. Generation and comprehension of unambiguous object descriptions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11–20, 2016.

Shuo Ni, Di Wang, He Chen, Haonan Guo, Ning Zhang, and Jing Zhang. UniGeoSeg: Towards unified open-world segmentation for geospatial scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 34865–34876, 2026.

Ruizhe Ou, Yuan Hu, Fan Zhang, Jiaxin Chen, and Yu Liu. GeoPix: A multimodal large language model for pixel-level image understanding in remote sensing. *IEEE Geoscience and Remote Sensing Magazine*, 13(3):324–337, 2025. doi: 10.1109/MGRS.2025.3560293.

Jerome Quenum, Wen-Han Hsieh, Tsung-Han Wu, Ritwik Gupta, Trevor Darrell, and David Chan. LISAt: Language-instructed segmentation assistant for satellite imagery. In *Advances in Neural Information Processing Systems*, volume 38, pp. 124660–124702, 2025. doi: 10.52202/085713-3756.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8748–8763. PMLR, 2021.

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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13009–13018, 2024.

Fu Rong, Meng Lan, Qian Zhang, and Lefei Zhang. RS2-SAM2: Customized SAM2 for referring remote sensing image segmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 8751–8759, 2026. doi: 10.1609/aaai.v40i11.37828.

Akashah Shabbir, Mohammed Zumri, Mohammed Bennamoun, Fahad Shahbaz Khan, and Salman Khan. GeoPixel: Pixel grounding large multimodal model in remote sensing. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 54095–54111. PMLR, 2025.

Hao Wang, Limeng Qiao, Zequn Jie, Zhijian Huang, Chengjian Feng, Qingfang Zheng, Lin Ma, Xiangyuan Lan, and Xiaodan Liang. X-SAM: From segment anything to any segmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 26187–26196, 2026a. doi: 10.1609/aaai.v40i31.39822.

Hao Wang, Limeng Qiao, Chi Zhang, Guanglu Wan, Lin Ma, Xiangyuan Lan, and Xiaodan Liang. X2SAM: Any segmentation in images and videos. In *European Conference on Computer Vision (ECCV)*, 2026b.

Junjue Wang, Zhuo Zheng, Ailong Ma, Xiaoyan Lu, and Yanfei Zhong. LoveDA: A remote sensing land-cover dataset for domain adaptive semantic segmentation. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, volume 1. Curran Associates, Inc., 2021.

Zhaoqing Wang, Yu Lu, Qiang Li, Xunqiang Tao, Yandong Guo, Mingming Gong, and Tongliang Liu. CRIS: CLIP-driven referring image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11686–11695, 2022.

Gui-Song Xia, Xiang Bai, Jian Ding, Zhen Zhu, Serge Belongie, Jiebo Luo, Mihai Datcu, Marcello Pelillo, and Liangpei Zhang. DOTA: A large-scale dataset for object detection in aerial images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3974–3983, 2018.

Zepeng Xin, Kaiyu Li, Luodi Chen, Wanchen Li, Xiao Yuchen, Hui Qiao, Weizhan Zhang, Deyu Meng, and Xiangyong Cao. SegEarth-R2: Towards comprehensive language-guided segmentation for remote sensing images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13199–13210, 2026.

Zhao Yang, Jiaqi Wang, Yansong Tang, Kai Chen, Hengshuang Zhao, and Philip H. S. Torr. LAVT: Language-aware vision transformer for referring image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18155–18165, 2022.

Zhigang Yang, Huiguang Yao, Linmao Tian, Xuezhi Zhao, Qiang Li, and Qi Wang. A large-scale referring remote sensing image segmentation dataset and benchmark. arXiv:2506.03583 [cs.CV], 2025.

Liang Yao, Fan Liu, Delong Chen, Chuanyi Zhang, Yijun Wang, Ziyun Chen, Wei Xu, Shimin Di, and Yuhui Zheng. RemoteSAM: Towards segment anything for earth observation. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pp. 3027–3036, 2025. doi: 10.1145/3746027.3754950.

Licheng Yu, Patrick Poirson, Shan Yang, Alexander C. Berg, and Tamara L. Berg. Modeling context in referring expressions. In *European Conference on Computer Vision (ECCV)*, pp. 69–85, 2016.

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.

Syed Waqas Zamir, Aditya Arora, Akshita Gupta, Salman Khan, Guolei Sun, Fahad Shahbaz Khan, Fan Zhu, Ling Shao, Gui-Song Xia, and Xiang Bai. iSAID: A large-scale dataset for instance segmentation in aerial images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pp. 28–37, 2019.

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.

Zilun Zhang, Zian Guan, Tiancheng Zhao, Haozhan Shen, Yuxiang Cai, Zhonggen Su, Yongheng Shang, Zhaojun Liu, Jianwei Yin, and Xiang Li. Geo-R1: Improving few-shot geospatial referring expression understanding with reinforcement fine-tuning. *ISPRS Journal of Photogrammetry and Remote Sensing*, 237:113–129, 2026. doi: 10.1016/j.isprsjprs.2026.04.023.

# APPENDIX

**Appendix Overview.** Appendices **A–C** specify the release format, evaluation splits, and query composition. Appendices **D–F** document the UAV imagery, public-source re-instrumentation, and public-track model evaluation. Appendices **G–K** detail annotation and human validation, vocabulary, difficulty construction, automated auditing, and diagnostic reference points. Appendices **L–M** provide model configurations, inference settings, uncertainty analysis, ranking reliability, and query-difficulty analyses. Appendices **N–O** collect qualitative error cases and analyses of paraphrase robustness and linguistic variation.

**A** Release layout and per-scene record . . . . . 15

**B** Released splits . . . . . 15

**C** Query composition by type, relation family and track . . . . . 15

**D** Source orthophotos and scene allocation . . . . . 16

**E** Public referring corpora . . . . . 16

**F** Public-track model evaluation . . . . . 17

**G** Annotation details . . . . . 19

**H** Vocabulary . . . . . 20

**I** Difficulty score . . . . . 20

**J** What the auditor found . . . . . 20

**K** Diagnostic reference points in full . . . . . 21

**L** Evaluation setup . . . . . 21

**M** Ranking reliability and what makes a query hard . . . . . 22

**N** Reporting atlas . . . . . 23

**O** Paraphrase robustness down the whole ladder . . . . . 23

## A RELEASE LAYOUT AND PER-SCENE RECORD

Each scene directory holds the image, a `uint16` instance index map, a `uint16` region bit-plane map, and a JSON record with: geodetic metadata (CRS, transform, ground sampling, scale, native pixel window); objects, each with category, kind, instance or region index, area in pixels and, where a ground sampling is known, in $m^2$, centroid in projected and geographic coordinates, bounding box; regions; relations as (subject, predicate, object) with metric distance where applicable; queries; and counterfactual pairs. Each query carries its text and two variants, its logical form, level, query type, relation family, track, target object ids, reference object ids, answer count, negativity and counterfactual flags, difficulty tier and score, candidate and distractor counts, answer area fraction, and the per-query agreement records of the relation-blind diagnostics.

## B RELEASED SPLITS

Table 5: Released splits. The first five protocols are the UAV track; the last two involve the public track. The unit is what is held out. *Buffer* scenes are used by no split; they exist so that a UAV train scene cannot share a pixel with a UAV test scene. `cross_scale` moves ground sampling and physical field of view together — 2048 pixels cover 512 m at one sampling and 256 m at the other — so it is a scale-shift protocol, not an intervention on resolution alone; no predicate changes meaning across it except the 5.2% of UAV queries flagged `window_calibrated` (Section 3), which a scale reading may exclude. `cross_corpus` trains on the UAV scenes and tests on the public ones. `source_official` is the assignment each public corpus published, carried through unchanged.

| protocol | unit | train | val | test | buffer | what it tests |
|---|---|---|---|---|---|---|
| `scene_disjoint` | scene | 447 | 70 | 140 | 43 | UAV, leakage-free |
| `geographic` | scene | 473 | – | 227 | – | urban → rural |
| `cross_scale` | scene | 101 | – | 249 | 350 | 0.25 m → 0.125 m |
| `reasoning_depth` | query | 15,617 | – | 5,299 | – | L1–L3 → L4–L5 |
| `compositional` | query | 17,851 | – | 3,065 | – | unseen (category, relation) |
| `cross_corpus` | scene | 700 | – | 1,010 | – | UAV → public |
| `source_official` | scene | 599 | 139 | 272 | – | inherited public splits |

## C QUERY COMPOSITION BY TYPE, RELATION FAMILY AND TRACK

The UAV collection contains 5,021 unanswerable queries (24.0%) and 2,477 counterfactual pairs over 4,577 queries; the public collection has 26.6% unanswerable queries and 3,698 pairs. Each logical form has three linguistic realizations. UAV queries have 29.2 same-category candidates on average (maximum 267), a median answer-area fraction of 0.33%, and mean sentence length of 7.8 words (13.2 at L5). The difficulty tiers contain 4,373 easy, 8,904 medium and 7,639 hard queries.

Table 6: Released UAV-track queries by generated type, relation family and evaluation track.

| type | queries | relation family | queries | track | queries |
|---|---|---|---|---|---|
| single relation | 6,052 | direction | 7,620 | robust | 8,250 |
| single attribute | 2,760 | logic | 3,646 | relational | 4,530 |
| multi relation | 2,739 | ordinal | 2,760 | compositional | 4,392 |
| negative relation | 2,273 | entity | 2,384 | basic | 3,065 |
| counterfactual | 2,134 | distance | 2,014 | set | 679 |
| multi hop | 1,653 | multi hop | 1,653 | | |
| negative object | 1,400 | topology | 839 | | |
| single entity | 984 | | | | |
| distance reasoning | 921 | | | | |

Tracks group UAV-track queries by what they test: basic naming, relational reference, compositional reference (L4–L5), robustness (unanswerable and counterfactual), and set-valued answers. A query belongs to exactly one track, so track scores partition the benchmark.

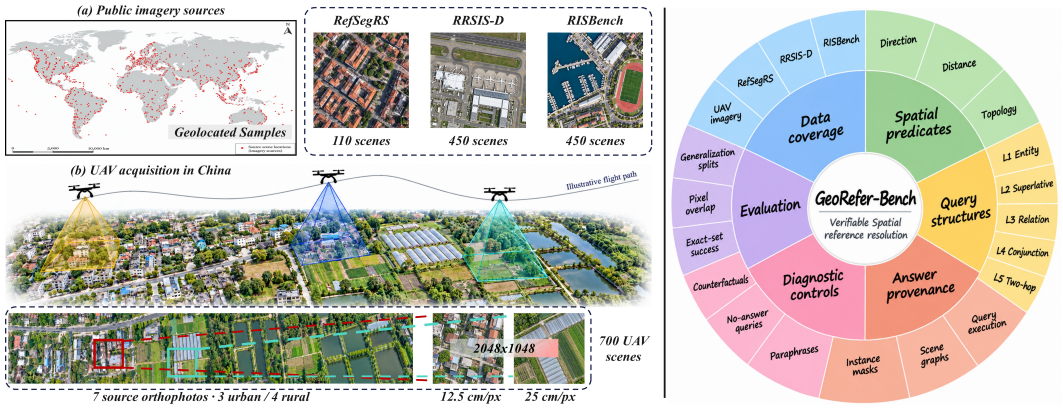


Figure 5: Data sources and benchmark overview. *Left:* public imagery from RefSegRS, RRSIS-D and RISBench, and seven Jiangxi UAV orthophotos (three urban and four rural), released as $2048^2$ scenes at 12.5 and 25 cm/px. *Right:* spatial predicates, L1–L5 query structures, answer provenance, diagnostic controls and evaluation read-outs. The two imagery tracks share the construction and verification workflow of Figure 2.

## D SOURCE ORTHOPHOTOS AND SCENE ALLOCATION

Figure 5 provides the data-source and benchmark overview that complements the construction workflow in Figure 2.

The source imagery covers 292.5 $\text{km}^2$ at 3.6–5.7 cm native sampling. Released scenes include 256 coarse windows (0.25 m/px, 512 m per side) and 444 fine windows (0.125 m/px, 256 m per side). Enforcing non-overlap throughout the collection would discard 55% of the available capacity; partially overlapping neighbours are retained subject to the primary split's native-pixel exclusion rule. UAV scene graphs average 203.99 edges per scene, with a maximum of 2,289.

Seven surveys, four rural and three urban, supply the UAV track; Table 7 gives each one's native ground sampling, area and scene count at both released samplings.

Table 7: Source orthophotos of the UAV track. Scenes are allocated by area, capped at 30% of each scale's budget, with shortfalls from exhausted orthos passed on; the largest survey therefore holds 38.9% of that track. `urban_c` is small enough to contribute a single scene and is retained for coverage, not for per-source comparison.

| source | domain | native GSD | area | coarse scenes | fine scenes |
|---|---|---|---|---|---|
| `urban_a` | urban | 5.63 cm | 213.0 $\text{km}^2$ | 123 | 149 |
| `urban_b` | urban | 5.53 cm | 37.4 $\text{km}^2$ | 66 | 134 |
| `urban_c` | urban | 5.42 cm | 0.5 $\text{km}^2$ | – | 1 |
| `rural_a` | rural | 5.42 cm | 7.5 $\text{km}^2$ | 17 | 29 |
| `rural_b` | rural | 5.65 cm | 14.6 $\text{km}^2$ | 30 | 58 |
| `rural_c` | rural | 3.64 cm | 15.2 $\text{km}^2$ | 19 | 65 |
| `rural_d` | rural | 5.45 cm | 4.3 $\text{km}^2$ | 1 | 8 |
| total | | | 292.5 $\text{km}^2$ | 256 | 444 |

## E PUBLIC REFERRING CORPORA

The public track is not a second benchmark. It is the same grammar, resolver and Exact Query Success applied to imagery this release did not choose, so a pattern measured on Jiangxi orthophotos can be checked on Vaihingen airborne tiles and on two web-map mosaics. Source sentences are retained as provenance and are not the queries we score. Table 8 is the inventory; Table 9 is the relation-blind diagnostic of Section 6 on those scenes; Table 10 compares the two protocols on identical pixels. Fifteen-model scores on the official test queries are Table 4 and Appendix F.

Table 8: Public track. Official splits are inherited. Metres are released only for RefSegRS, whose ground sampling we calibrate from its own vehicle masks rather than taking the documented 0.13 m/px on trust (below). RRSIS-D and RISBench do not publish a per-image sampling, so their scenes live in the image plane and the metric distance family is withheld.

| source | imagery | px | scenes | instances | queries | units |
|---|---|---|---|---|---|---|
| RefSegRS | ISPRS Vaihingen | 512 | 110 | 5,418 | 3,278 | metre (0.30) |
| RRSIS-D | DIOR / Google Earth | 800 | 450 | 8,981 | 12,774 | pixel |
| RISBench | Google Earth | 512 | 450 | 9,872 | 12,544 | pixel |
| total | | | 1,010 | 24,271 | 28,596 | |

**What we did not take on trust.** RefSegRS documents 0.13 m/px. In its own vehicle masks a passenger car covers a median of 88 px and spans a median of 15 px, which at 0.13 m would be a 1.5 $m^2$ car. Three estimators over those 36,076 car footprints — footprint area against a $4.5 \times 1.85$ m sedan, longest bounding-box side against 4.5 m, shortest against 1.85 m — give 0.308, 0.300 and 0.264 m/px; 0.30 is used here, and it is also what a 1200 px tile at 0.13 m becomes when resampled to 512 px. Under the documented figure every area in these scenes would have been wrong by 5.5× and every distance by 2.3×.

RISBench's validation set is declared only by which phrase file an entry appears in, and all 10,013 validation entries are named `train_...`. Reading the split off the filename prefix would put 5,561 validation images into training. The declaration is used instead; no image has its targets spread across two phrase files, so the assignment is well defined per scene.

**Independent masks.** Each source already drew a mask for the object its own sentence refers to. Those masks were not used to build our instances, so they remain an independent check. Median IoU against our inventory: 0.35 on RefSegRS (class-union masks; 0.81 on buildings, 0.38 on vehicles), 0.58 on RRSIS-D and 0.67 on RISBench (instance masks). Near-duplicate imagery across the three sources: none at a 64-bit difference hash with Hamming distance at most 4. The degree to which relational queries require selection beyond category membership is measured by the all-category diagnostic in Table 9.

**Categories.** Discrete instances on the public track: building 13,577, vehicle 7,322, parking lot 756, overpass 499, ship 414, pond 389, tennis court 343, storage tank 259, baseball field 200, airplane 141, chimney 93, basketball court 85, stadium 69, running track 57, bridge 47, windmill 8, golf course 6, dam 6. Golf course and dam are thin: the annotator answers those prompts rarely even on scenes whose source label is one of them, and they are released with their counts stated rather than dropped. Region anchors additionally include farmland and runway.

Table 9: Relation-blind selection on the public track, in per cent. Per-level columns are over answerable queries. Naming the category still scores 100 at L1 and 0 at L2, L4 and L5; overlap still pays the same predictions 37.4 mIoU. Compare Table 13.

| strategy | all 28,596 queries | | per level, answerable | | | | |
|---|---|---|---|---|---|---|---|
| | EQS | mIoU | L1 | L2 | L3 | L4 | L5 |
| all of category | 20.2 | 37.4 | 100.0 | 0.0 | 10.0 | 0.0 | 0.0 |
| largest instance | 26.2 | 34.1 | 70.0 | 32.9 | 17.2 | 8.8 | 21.3 |
| random instance | 23.0 | 26.6 | 70.0 | 17.5 | 16.5 | 6.5 | 15.4 |

## F PUBLIC-TRACK MODEL EVALUATION

The UAV ranking in Table 3 is one imagery regime. The public track re-instruments RefSegRS, RRSIS-D and RISBench under the same grammar, so the same fifteen systems can be scored on imagery this release did not collect; Table 4 in the main text reports them. Table 11 is the denominator: each source's official test assignment, 272 scenes and 7,879 canonical queries in total — not the 1,010/28,596 inventory, and not the UAV 1,881-query sample. Scores are not pooled with Table 3, and they are not a UAV-to-public transfer experiment (that split remains the `cross_corpus` row of Table 5).

Table 10: Two protocols on the same pixels. *Source* is the original expression of RefSegRS, RRSIS-D or RISBench for that photograph; *ours* is a query generated from the scene graph of that photograph. Relational share is the fraction of sentences that commit to a spatial relation; unanswerable is a sentence whose correct output is nothing.

| corpus | protocol | $n$ | multi-obj. | area% | words | relational | unans. |
|---|---|---|---|---|---|---|---|
| RefSegRS | source | 1,772 | 0.00 | 1.36 | 2.0 | 0.16 | 0.00 |
| | ours | 3,278 | 0.34 | 0.27 | 7.0 | 0.75 | 0.23 |
| RRSIS-D | source | 450 | 0.00 | 2.04 | 6.0 | 0.00 | 0.00 |
| | ours | 12,774 | 0.23 | 0.66 | 7.0 | 0.71 | 0.27 |
| RISBench | source | 785 | 0.00 | 0.89 | 14.0 | 0.27 | 0.00 |
| | ours | 12,544 | 0.25 | 0.55 | 7.0 | 0.70 | 0.27 |

**Protocol.** All fifteen comparison models are evaluated using their respective model checkpoints and inference pipelines, documented in Appendix L, with the common read-out of Equation 3. The input is the published patch at native size (512 or 800), the text is the canonical wording of a GeoRefer-Bench query, and the logical form is never supplied. Source-corpus expressions are provenance and are not scored. $\text{EQS}_a$ is exact query success over every official-test query of that source; $\text{EQS}_+$ and mIOU are restricted to answerable queries; L1–L5 are answerable EQS at that level, the level columns of Table 4. Per-source rejection accuracy and counterfactual-pair consistency are part of the released per-query records rather than columns here. All these tables use the canonical wording, as Table 3 does.

**What to read.** Table 12 asks whether the family ordering survives the change of imagery; the per-source level blocks of Table 4 expose the depth profile inside each source. RefSegRS L1 and L5 rest on 50 and 47 answerable queries; those two cells are thin relative to RRSIS-D and RISBench. A depth drop on all three sources is evidence that the UAV pattern is not an artefact of Jiangxi orthophotos under this grammar. It is not evidence about unrestricted human language, and it does not rank the fifteen systems on a mixture of the two tracks.

Table 11: Public-track inventory against the official test subset actually scored. Inventory counts repeat Table 8; test counts are the inherited assignment.

| | Complete public-track inventory | | | Official test | |
|---|---|---|---|---|---|
| Source | Scenes | Objects | Queries | Scenes | Queries |
| RefSegRS | 110 | 5,418 | 3,278 | 40 | 1,196 |
| RRSIS-D | 450 | 8,981 | 12,774 | 90 | 2,461 |
| RISBench | 450 | 9,872 | 12,544 | 142 | 4,222 |
| Total | 1,010 | 24,271 | 28,596 | 272 | 7,879 |

Table 12: Fifteen models on the public-track official test queries, in per cent, per source. $EQS_a$ is all-query EQS; $EQS_+$ is answerable-query EQS; IoU is answerable mIoU. Macro is the equal-source mean of $EQS_a$ and is not weighted by query count. Best per column in **red**, runner-up underlined. Row order matches Table 3.

| | RefSegRS | | | RRSIS-D | | | RISBench | | | Macro |
|---|---|---|---|---|---|---|---|---|---|---|
| Method | $EQS_a$ | $EQS_+$ | IoU | $EQS_a$ | $EQS_+$ | IoU | $EQS_a$ | $EQS_+$ | IoU | $EQS_a$ |
| *Remote-sensing reasoning segmentation* | | | | | | | | | | |
| SegEarth-R2 | 73.1 | 75.6 | **67.9** | **74.6** | **77.9** | **66.8** | 72.3 | 76.6 | **69.2** | 73.3 |
| Geo-R1 | **74.7** | **77.5** | 67.2 | 73.6 | 77.2 | 62.9 | **73.2** | **77.8** | 66.6 | **73.8** |
| UniGeoSeg | 72.2 | 74.3 | 65.2 | 71.4 | 75.3 | 61.1 | 69.5 | 73.4 | 63.8 | 71.0 |
| LISAt | 59.9 | 62.4 | 48.8 | 62.6 | 65.7 | 46.3 | 60.5 | 64.1 | 48.8 | 61.0 |
| SegEarth-R1 | 45.3 | 44.8 | 30.7 | 47.7 | 49.5 | 31.4 | 49.8 | 51.9 | 35.1 | 47.6 |
| *Remote-sensing foundation and grounding models* | | | | | | | | | | |
| GeoPixel | 59.6 | 62.5 | 47.1 | 62.8 | 66.9 | 45.6 | 61.5 | 64.9 | 47.1 | 61.3 |
| RemoteSAM | 39.9 | 40.1 | 31.3 | 44.0 | 46.0 | 33.4 | 42.6 | 44.6 | 33.6 | 42.2 |
| GeoPix | 40.2 | 38.9 | 24.7 | 44.8 | 47.0 | 28.2 | 43.9 | 45.7 | 28.0 | 43.0 |
| *Classical RRSIS* | | | | | | | | | | |
| RS2-SAM2 | 37.0 | 35.6 | 31.2 | 42.3 | 44.3 | 36.6 | 40.1 | 42.1 | 36.7 | 39.8 |
| RSRefSeg2 | 33.5 | 32.4 | 26.6 | 39.6 | 41.9 | 32.2 | 37.9 | 39.5 | 32.3 | 37.0 |
| FIANet | 24.7 | 22.7 | 19.7 | 31.0 | 31.7 | 27.4 | 30.6 | 31.7 | 28.2 | 28.8 |
| CADFormer | 23.2 | 21.0 | 18.8 | 29.5 | 29.8 | 25.1 | 28.8 | 29.5 | 25.9 | 27.2 |
| RMSIN | 23.2 | 20.8 | 17.7 | 26.7 | 26.3 | 21.6 | 27.2 | 27.2 | 23.7 | 25.7 |
| *Generalist segmentation MLLM* | | | | | | | | | | |
| X2SAM | 28.4 | 25.4 | 20.8 | 33.4 | 32.2 | 24.0 | 33.2 | 32.5 | 25.9 | 31.7 |
| X-SAM | 26.0 | 23.0 | 18.7 | 31.5 | 30.8 | 22.7 | 31.6 | 31.5 | 24.5 | 29.7 |

## G ANNOTATION DETAILS

Proposals from the two branches are merged by dropping anything clipped by an interior tile edge, score-ordered non-maximum suppression on bounding-box-cropped masks, one category per blob, and a final pass that makes instances pixel-disjoint, with contested pixels going to the smaller object and geometry re-measured from the result.

Annotation uses model assistance with domain-expert verification at the category, scene, boundary and query levels. Candidate categories and proposal thresholds are reviewed on pilot scenes at both UAV ground samplings. SAM3-based proposals supply an initial object inventory, while boundary evidence and LLM-assisted checks support inspection of its visual and semantic consistency. Experts repeatedly verify the annotations against the imagery. Completeness is therefore assessed through the combined workflow rather than inferred from the segmenter's proposal count or confidence alone.

**Human annotation validity.** Expert validation targets three sources of benchmark error separately: object inventory and mask quality, the semantic validity of spatial references, and consistency between expressions and their resolved answer sets. Reviewers inspect whether annotated objects are visually supported, whether each retained expression admits the intended spatial interpretation in the scene, and whether the resulting answer set agrees with that interpretation. A human-traced subset additionally provides an independent check of boundary quality rather than relying only on promptable segmentation proposals. These checks complement executable validation: program execution verifies internal consistency, whereas human review addresses errors that cannot be detected from the scene graph alone, such as missed objects, ambiguous references, or visually unsupported relations.

The reviewed image and instance masks, together with the L1–L5 specification, condition LLM-assisted proposals of expressions, logical forms and candidate object IDs. Domain experts review anchor identifiability, spatial validity and expression–program correspondence; ambiguous or unsupported candidates are corrected or excluded. The resolver then computes the answer set from the accepted form and scene records, and the query mask is the union of the selected instance masks. Other pixels are background in that query mask; the complete scene inventory is retained. Experts inspect the resulting expression, answer IDs and mask together.

Domain experts also repeatedly review spatial reference resolution: whether the expression, its anchor and spatial constraints agree with the intended target set in the scene. This human check complements the resolver, which establishes the output of a formal program but cannot by itself establish that the wording has the intended human interpretation. Repeated expert review is reported as a validation procedure, separately from any quantitative inter-annotator agreement statistic.

A subset of object boundaries is manually traced and repeatedly checked by domain experts. It provides a human-delineated reference for inspecting boundary quality beyond the proposal masks. This subset is distinct from scene-level inspection and from comparison with existing public-source masks. Its presence does not imply that every released mask was hand-traced or that boundary fidelity alone establishes inventory completeness. The quality assessment distinguishes three questions: whether relevant objects are represented, whether their masks follow the visible boundaries, and whether a spatial expression resolves to the intended objects.

Carrying an unreliable category does not add coverage, it corrupts the relation graph. One category (sports field, which the segmenter returned for rectangular farm plots) was removed only after a pilot produced "the greenhouses east of the sports field" in a scene with six of them. Anchor selection also skips trivial containment pairs (pond–water, building–residential), since the region already contains those instances by construction, and a category is never its own reference.

## H VOCABULARY

Discrete categories on the UAV track, with released instance counts: building $23{,}447$, pond $2{,}340$, greenhouse $201$, parking lot $200$, storage tank $18$, ship $9$, bridge $2$. Region anchors: road, water, forest, farmland, residential; $3.59$ per scene on average. Rejected on that track with recorded reasons: vehicle (too small at these ground samplings), sports field (returned for rectangular farm plots), shrub, grassland, bare soil, sand, plaza, commercial, harbor, runway, rice paddy, industrial, mountain, solar panel, wetland. Public-track categories are listed in Appendix E.

## I DIFFICULTY SCORE

For a query with $c$ same-category candidates, level $\ell$, answer area fraction $f$ and answer size $k$:

$$d \;=\; \min\!\big(3,\tfrac{c-1}{6}\big) \;+\; 0.6\,(\ell-1) \;+\; 1.5\cdot\mathbb{1}[f<10^{-3}] \\ +\; 0.75\cdot\mathbb{1}[10^{-3}\le f<5\times10^{-3}] \;+\; 0.5\cdot\mathbb{1}[k>1] \;+\; 0.5\cdot\mathbb{1}[k>3],$$

with tiers easy ($d<2$), medium ($2\le d<4$) and hard ($d\ge 4$). Every term is a property of the scene or the sentence, so the tiers do not change when a baseline does.

## J WHAT THE AUDITOR FOUND

*Objects cut by tile seams.* $8.8\%$ of instances had a bounding-box edge within $2$ pixels of the tile boundary at $x=1024$, against $1\%$ expected if edges were spread uniformly, because contiguous tiles cut objects and both halves survived non-maximum suppression by overlapping the whole object about half. Tiles now overlap by $512$ pixels and any proposal touching an interior tile edge is dropped, since some other pass sees that object whole. After the fix, $0$ of $182$ sampled instances lie on the seam and the tile-edge density ratio fell from $24$–$70\times$ to $0.3$–$1.9\times$.

*Records disagreeing with their own masks.* With overlapping proposals written into one index map, five instances per scene had no pixels left and eight had bounding boxes the mask no longer supported. Instances are now made pixel-disjoint before anything is written, contested pixels going to the smaller object, and all geometry is re-measured from the result, so the map and the metadata agree by construction.

*Ambiguous anchors.* Generation used the largest instance of a reference category regardless of how many existed, producing sentences such as "the ponds within $92$ metres of the pond" in a scene holding $32$ ponds. The scene builder had already computed which anchors were nameable and was simply not passing them to the generator; distance queries had hidden the bug because they fall back to centroid distance when no edge exists. Nameability is now a precondition of generation, not a filter applied afterwards.

*Concurrent builders.* Two builders once ran into one release tree and each deleted scenes the other wrote; of $760$ scenes in that tree, only $96$ were provably clean. The tree was discarded, the builder now takes an exclusive lock on the release directory, and the provenance audit that found this is part of the release.

## K DIAGNOSTIC REFERENCE POINTS IN FULL

Table 13 is the long form of the right-hand block of Table 2: every relation-blind diagnostic of Section 6 under both criteria on the UAV track, with the bounds that frame the scale. The public-track counterpart is Table 9.

Table 13: Diagnostic reference points on the UAV track, in per cent, with what each one exploits named under its row. Per-level columns are over answerable queries, exact-set EQS above and overlap mIOU in grey below; the arrow says which way overlap moves against the exact-set number over it, so a red ↑ is overlap paying for a prediction the set criterion rejects. Within the relation-blind block, best per column in **red** and runner-up underlined.

| strategy | all 20,916 queries | | per level, answerable: EQS above, mIOU below | | | | |
|---|---|---|---|---|---|---|---|
| | EQS | mIOU | L1 | L2 | L3 | L4 | L5 |
| *Relation-blind selection, given ground-truth masks* | | | | | | | |
| all of category | 16.0 | 29.7 | 100.0 | 0.0 | 9.2 | 0.0 | 0.0 |
| names the category | | | 100.0 = | 12.2 ↑ | 36.9 ↑ | 11.8 ↑ | 13.3 ↑ |
| largest instance | 22.7 | 28.6 | 71.5 | 30.4 | 16.9 | 5.5 | 13.4 |
| a size prior | | | 69.4 ↓ | 30.4 = | 32.0 ↑ | 11.8 ↑ | 13.4 = |
| random instance | 19.0 | 20.3 | 71.5 | 10.1 | 15.6 | 3.6 | 13.6 |
| chance among candidates | | | 52.1 ↓ | 10.1 = | 25.8 ↑ | 7.3 ↑ | 13.6 = |
| *Bounds, by construction* | | | | | | | |
| always silent returns nothing | 24.0 | – | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| the answer set oracle | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 |

## L EVALUATION SETUP

Table 3 and the plates of Section 7 and Appendix N report predictions obtained by running all fifteen comparison models with their respective released or officially provided checkpoints and inference pipelines. The four families are remote-sensing reasoning segmentation — SegEarth-R2 (Xin et al., 2026), Geo-R1 (Zhang et al., 2026), UniGeoSeg (Ni et al., 2026), LISAt (Quenum et al., 2025) and SegEarth-R1 (Li et al., 2025b) —, remote-sensing foundation and grounding models — GeoPixel (Shabbir et al., 2025), RemoteSAM (Yao et al., 2025) and GeoPix (Ou et al., 2025), two pixel-grounding multimodal language models and one segmentation foundation model —, classical referring remote-sensing image segmentation — RS2-SAM2 (Rong et al., 2026), RSRefSeg2 (Chen et al., 2026), FIANet (Lei et al., 2025), CADFormer (Liu et al., 2025) and RMSIN (Liu et al., 2024) — and generalist segmentation MLLMs — X2SAM (Wang et al., 2026b) and X-SAM (Wang et al., 2026a). The protocol is identical for all of them: the input is the whole $2048^2$ scene, the text is the released sentence, the logical form is never supplied, and $\mathcal{G}$ is offered as optional metadata that a system consumes only where its own interface accepts such fields — most of them take an image and a string, so for those the geodetic record is unused. Returned pixels are read back as an instance set by Equation 3, so a system that emits one mask and a system that emits many are scored by the same criterion. Where an official interface could not accept a full scene, preprocessing was adapted to its input requirements while preserving the evaluated model weights and decoding procedure. Checkpoint identifiers and the corresponding prompt formats, thresholds, tiling and post-processing settings are documented in the released evaluation configurations.

Evaluation runs over a fixed sample of 130 test scenes from the UAV scene-disjoint split, with up to three answerable queries per reasoning level and three unanswerable queries per scene — 1,881 queries in total. Predictions are stored per (model, query id), so a cell in one plate is the same prediction as the corresponding cell in another and every plate redraws identically. Predictions from the comparison runs are not used to define ground truth: answer membership comes from the reviewed scene records and Equation 1, and answer pixels are the full-resolution instance masks.

Appendix F evaluates all fifteen comparison models on every canonical query in the public official test assignment: 40 RefSegRS scenes (1,196 queries), 90 RRSIS-D scenes (2,461) and 142 RISBench scenes (4,222). Public patches are read at native size. Those tables do not reuse the UAV 1,881-query sample and do not mix the two tracks.

## M RANKING RELIABILITY AND WHAT MAKES A QUERY HARD

**How much of the ordering is real.** Queries within one scene share objects, imagery and often an anchor, so the resampling unit is the scene: we draw 130 scenes with replacement, 10,000 times, and recompute every system's overall EQS on the same resample, which makes the comparison paired. The 95% interval is $\pm 1.8$ to $\pm 2.3$ points wide for every system (SegEarth-R2 74.1, $[72.0, 76.1]$; Geo-R1 73.2, $[71.2, 75.1]$; UniGeoSeg 70.1, $[67.9, 72.3]$; LISAt 59.6, $[57.3, 61.8]$; GeoPixel 59.5, $[57.2, 61.8]$; SegEarth-R1 44.7, $[42.5, 47.0]$; RemoteSAM 39.9, $[37.9, 42.0]$; GeoPix 39.9, $[37.8, 42.0]$; RS2-SAM2 38.3, $[36.3, 40.4]$; RSRefSeg2 37.1, $[35.4, 39.0]$; X2SAM 32.4, $[30.6, 34.3]$; X-SAM 29.9, $[28.0, 31.8]$; FIANet 29.7, $[27.9, 31.7]$; CADFormer 27.7, $[25.8, 29.6]$; RMSIN 25.5, $[23.8, 27.4]$). Taken one at a time, 98 of the 105 pairs are separated at 0.05; the largest difference the paired resample cannot distinguish from zero is 1.5 points (RemoteSAM against RS2-SAM2) and the smallest it does distinguish is 2.0 (FIANet against CADFormer). But a leaderboard is read as 105 comparisons, not one, so we also adjust the same bootstrap $p$-values by the Holm step-down procedure, which controls the probability of *any* false separation: 89 pairs survive, the largest unseparated gap widens to 3.9 points (SegEarth-R2 against UniGeoSeg) and the smallest separated one to 4.2. Four points is therefore the honest resolution of the ordering. The top three systems are one band, and so are FIANet, X-SAM, CADFormer and RMSIN at the bottom; the statements the paper makes — the lead of the reasoning family, the 25.5–38.3 floor of the other two, the $\Delta_{1\to5}$ split — are 20 points or more and are unaffected.

**What the queries that fail have in common.** Table 15 bins the same per-query records three ways. Same-category candidates: at L3, where the logical form is fixed at one relation to a named anchor, so depth cannot confound the comparison, the field mean falls from 49.0 with at most five candidates to 37.1, 34.3 and 34.1 above fifty, and the systems that hold up are the ones that hold up on the benchmark as a whole. Answer area: splitting the same answerable queries at the median moves the field mean in the *opposite* direction, 42.5 for the smaller half against 47.8 for the larger, indicating lower success for smaller answer regions. Sentence length: short sentences score 61.3 and long ones 26.2, the largest of the three effects, but length here covaries with composition and is not established as an independent cause — L4 and L5 forms carry two relations and therefore more words — which is why depth and ambiguity are reported as separate columns of Table 3 and not merged into one difficulty score.

Candidate count and answer area are correlated in overhead imagery, since the categories that appear in dozens are also the small ones, so the two are separated by crossing them inside one level. Within L3, split into terciles of answer area, the field mean at most five candidates against more than five is 54.1 against 36.6 for the smallest tercile, 46.8 against 34.5 for the middle and 45.4 against 35.6 for the largest: a 9.7 to 17.4 point effect of ambiguity in *every* area stratum, in the same direction, while area itself moves the score by at most 8.7 points at fixed ambiguity. The association with candidate ambiguity persists within these area strata; this does not rule out a contribution from object size or other correlated scene properties.

**Two things the ordering does not depend on.** Table 14 re-scores the same predictions under two interventions on the query population. *Category balance*: buildings are 53.7% of the released queries and 1,091 of the 1,881 test items, so the leaderboard could be a building leaderboard. Averaging over the two target categories that carry at least 40 test queries instead of over queries (building, pond) raises every system by 2 to 6 points and exchanges two adjacent pairs, RS2-SAM2 with RSRefSeg2 and FIANet with X-SAM, both inside the 4-point band; Kendall's $\tau$ against the reported ordering is 0.96 and no system moves more than one place. The two categories with enough support are not the whole vocabulary, which is a limit on what can be said per category, not on the ordering. *Scale-dependent semantics*: 54 of the 1,881 test queries (2.9%) use *near* or *far*, whose metres are calibrated to the scene's ground window. Removing them changes no system by more than 0.5 points and no rank at all ($\tau = 0.99$), so nothing in Table 3 rests on the one part of the predicate vocabulary whose physical meaning depends on how the scene was sampled.

Table 14: The leaderboard re-scored under two interventions on the query population: averaging over target categories rather than over queries, and dropping the queries whose distance word is calibrated to the ground window. EQS over all queries, with the rank each scoring implies.

| method | as reported | | category-balanced | | scale-invariant only | |
|---|---|---|---|---|---|---|
| | EQS | rank | EQS | rank | EQS | rank |
| SegEarth-R2 | 74.1 | 1 | 79.9 | 1 | 73.7 | 1 |
| Geo-R1 | 73.2 | 2 | 76.6 | 2 | 73.2 | 2 |
| UniGeoSeg | 70.1 | 3 | 74.4 | 3 | 70.1 | 3 |
| GeoPixel | 59.5 | 5 | 62.0 | 5 | 59.0 | 5 |
| LISAt | 59.6 | 4 | 62.1 | 4 | 59.3 | 4 |
| SegEarth-R1 | 44.7 | 6 | 47.6 | 6 | 44.7 | 6 |
| RemoteSAM | 39.9 | 7 | 42.3 | 7 | 39.7 | 7 |
| GeoPix | 39.9 | 8 | 41.9 | 8 | 39.7 | 8 |
| RS2-SAM2 | 38.3 | 9 | 40.6 | 10 | 37.9 | 9 |
| RSRefSeg2 | 37.1 | 10 | 41.3 | 9 | 36.9 | 10 |
| X2SAM | 32.4 | 11 | 33.3 | 11 | 32.6 | 11 |
| FIANet | 29.7 | 13 | 32.9 | 12 | 29.9 | 13 |
| X-SAM | 29.9 | 12 | 31.7 | 13 | 30.1 | 12 |
| CADFormer | 27.7 | 14 | 30.1 | 14 | 27.9 | 14 |
| RMSIN | 25.5 | 15 | 28.6 | 15 | 25.7 | 15 |

Table 15: What makes a query hard, in EQS over answerable queries: same-category candidates, answer area against the median, and sentence length against the median, for six representative systems spanning the four model families and for the mean over all fifteen. These descriptive associations motivate the within-level, answer-area-stratified analysis; they do not isolate causal effects.

| method | same-category candidates | | | | answer area | | sentence length | |
|---|---|---|---|---|---|---|---|---|
| | 1–5 | 6–20 | 21–50 | >50 | small | large | short | long |
| SegEarth-R2 | 82.6 | 72.1 | 71.2 | 77.9 | 76.0 | 78.0 | 85.7 | 66.8 |
| Geo-R1 | 82.4 | 71.2 | 71.5 | 69.7 | 72.9 | 77.4 | 83.2 | 65.8 |
| UniGeoSeg | 79.9 | 73.7 | 63.5 | 66.6 | 71.8 | 73.7 | 84.0 | 59.7 |
| GeoPixel | 72.1 | 57.8 | 53.3 | 48.3 | 59.2 | 61.7 | 73.9 | 44.8 |
| RMSIN | 40.9 | 15.9 | 12.0 | 10.7 | 17.7 | 29.5 | 38.3 | 6.4 |
| X-SAM | 42.0 | 21.4 | 19.0 | 13.4 | 23.9 | 30.5 | 45.6 | 5.5 |
| **field mean** | 57.3 | 40.2 | 36.9 | 35.5 | 42.5 | 47.8 | 61.3 | 26.2 |

## N REPORTING ATLAS

The plates below are the rest of the evaluation, in the order a results section would use them: the leaderboard layout (Figure 7), qualitative galleries at both ground samplings (Figures 8–10), the error taxonomy (Figure 11), unanswerable queries (Figure 12), consensus maps (Figure 13), the depth ladder on other scenes and at the coarse sampling (Figures 14–16), paraphrase stability (Figure 17), and two further quantitative breakdowns (Figures 18 and 19). Appendix O then crosses rewording with depth over the whole ladder. All of them are drawn from the stored predictions of the fifteen evaluated systems (Appendix L).

## O PARAPHRASE ROBUSTNESS DOWN THE WHOLE LADDER

Figure 17 states the rewording format on two levels, which is enough to define it and not enough to answer the question it exists for: whether a method's stability under rewording survives the sentence getting deeper. Figures 20–23 are that measurement in full, one plate per scene, one query at each of the five levels, three surface forms each. They read in three directions. Across a row is the usual method comparison. Down a block is one meaning said three ways, so a block whose cells change is a column resolving the template rather than the geometry. Down the plate is composition depth, which is where the interesting pattern sits: a column can be perfectly stable at L1 and L2 — where the sentence has one clause and rewording it changes nothing — and lose that stability from L3 on, once a relation has to be parsed out of a rephrased clause. Four scenes are shown so that the pattern is

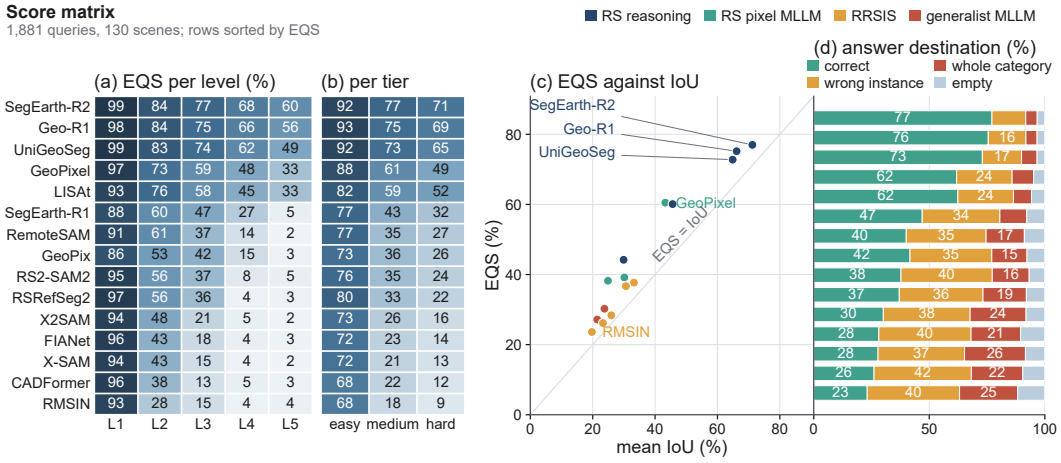


Figure 6: The quantitative format: score matrices per reasoning level and per difficulty tier, the EQS–mIOU plane, and where wrong answers go. The plane is the argument of Section 6 drawn per method — a system can sit far to the right of the diagonal, overlapping the answer, while returning the wrong set.

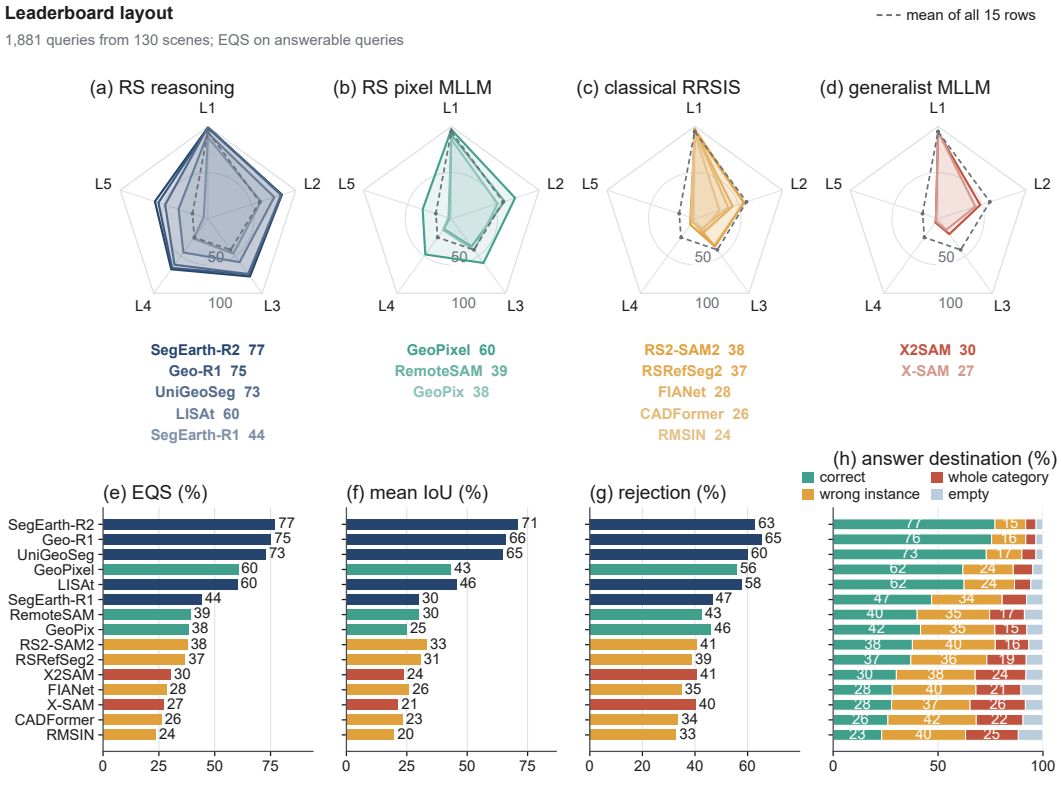


Figure 7: The leaderboard layout of Table 3 as a figure. Radars give exact query success per reasoning level, grouped by method family so that a family's shape — flat, or collapsing after L3 — is visible before any single number is; bars give the four quantities a submission must report.

not a property of one image: two urban and one rural footprint at $12.5$ cm, and one $512$ m footprint at $25$ cm. Each plate carries seven of the fifteen columns, two per family, because fifteen rows and fifteen columns on one page would put the labels below $7$ pt.

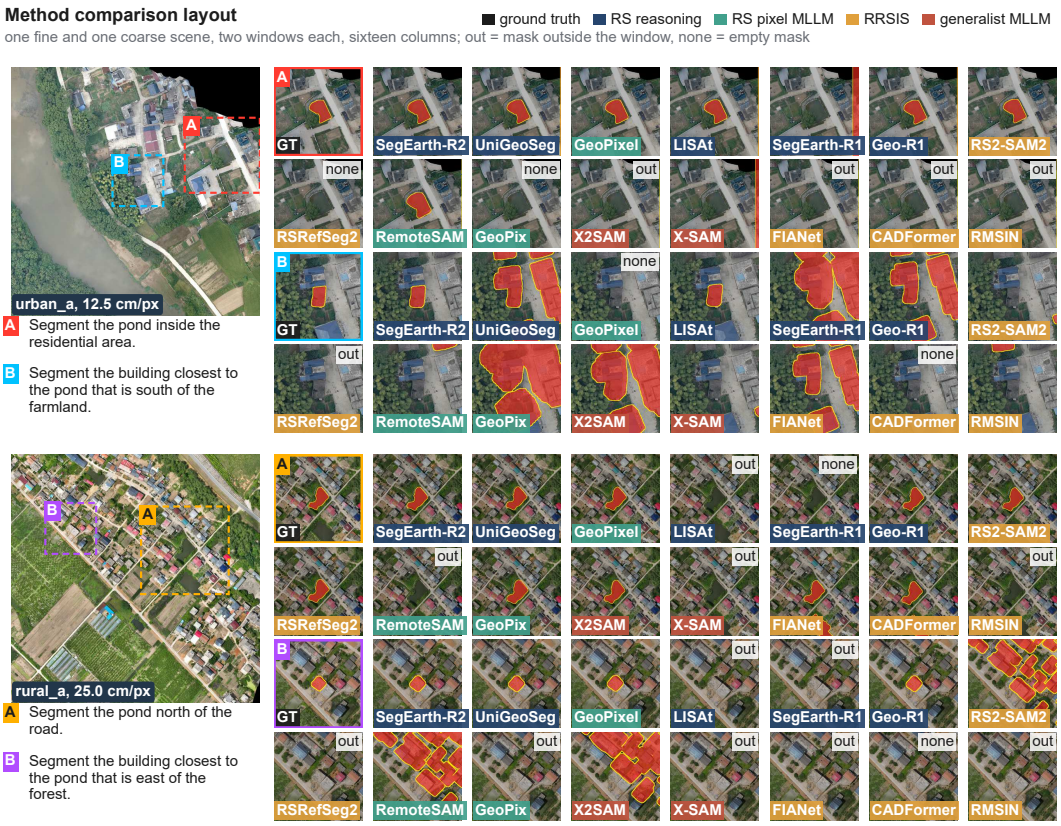


Figure 8: Figure 4 on further scenes: the whole scene and its two sentences on the left, then the windows they refer into across the ground truth and every evaluated model.




Figure 9: The same layout on the rural sources at 25 cm ground sampling, where answers are a smaller fraction of the scene and anchors are regions rather than objects.

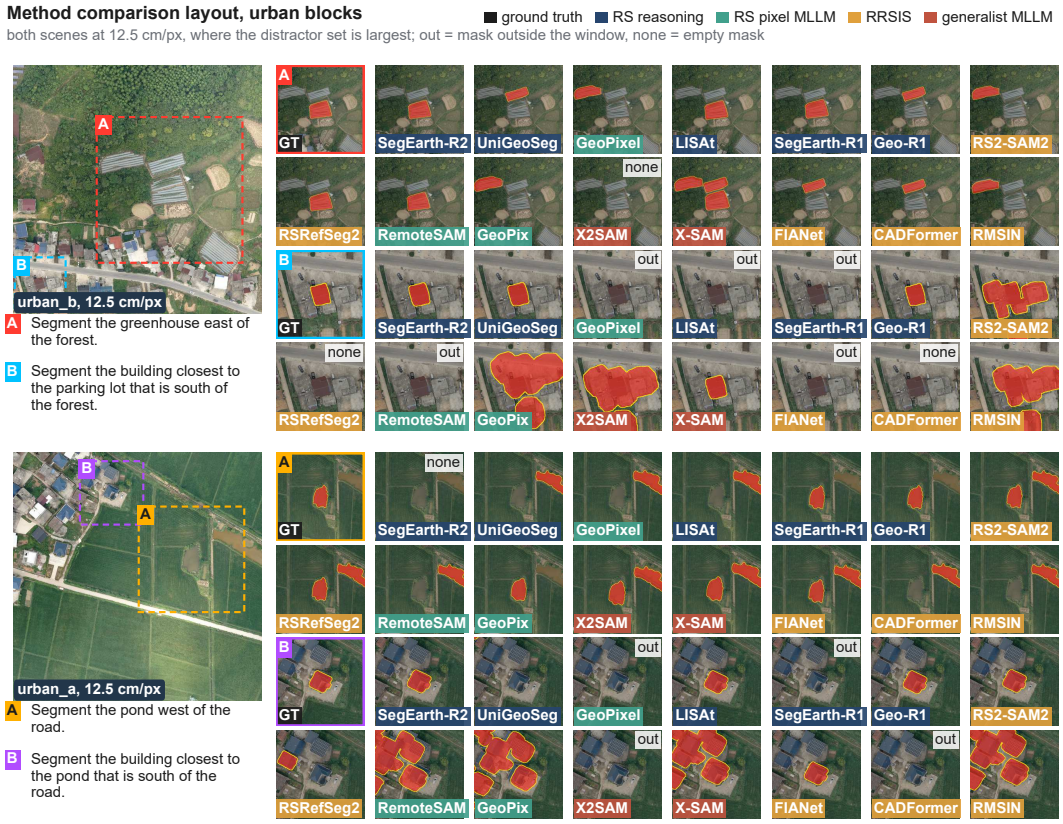


Figure 10: The comparison layout on dense urban blocks at 12.5 cm, the setting with the most same-category candidates per query.

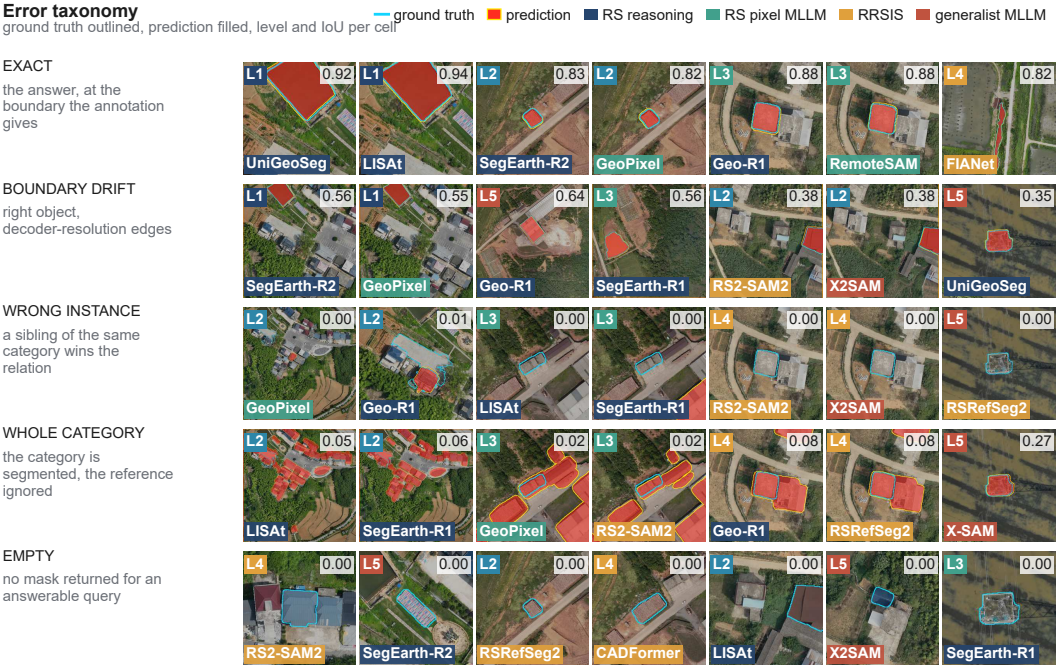


Figure 11: The error taxonomy a submission is scored against: exact, boundary drift, a sibling instance, the whole category, and an empty mask. All reported EQS uses the read-out of Equation 3, under which the first two read back as the same correct object set and both earn EQS — boundary drift is what mIoU penalises and EQS does not, which is why overlap is reported alongside it — while the last three return the wrong object set and fail. Under the native-instance matching of Equation 2, a drifted mask also fails EQS once its IoU falls below $\tau$.




Figure 12: Unanswerable queries at full scene width. The correct output is nothing at all; a returned mask is a hallucination and is scored as one, which is why rejection accuracy is reported beside EQS.




Figure 13: Consensus maps: every method's prediction summed per pixel, against the answer. Where the consensus is confident and wrong, the query is a systematic trap rather than a hard instance.

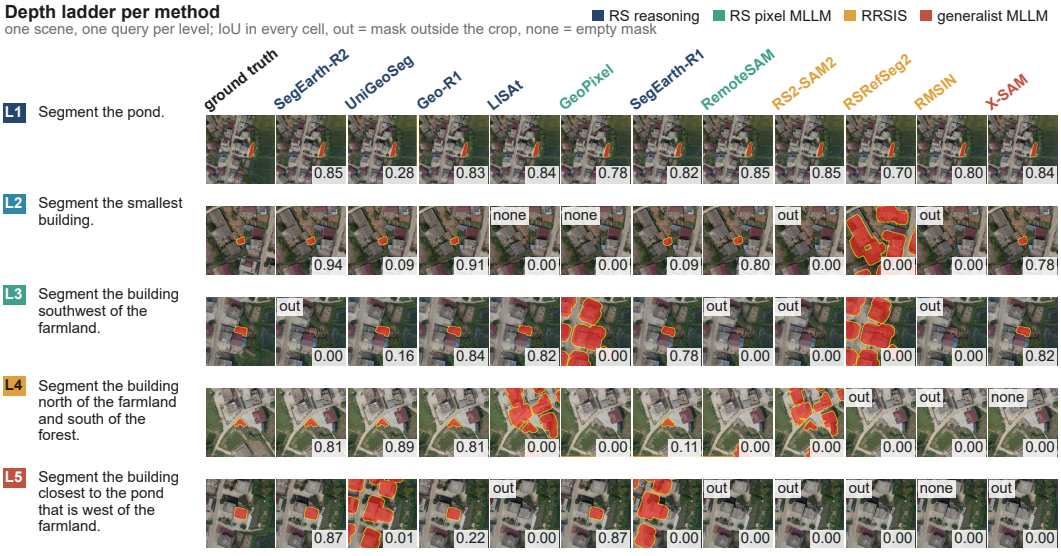


Figure 14: The depth ladder of Figure 3 with all fifteen columns.

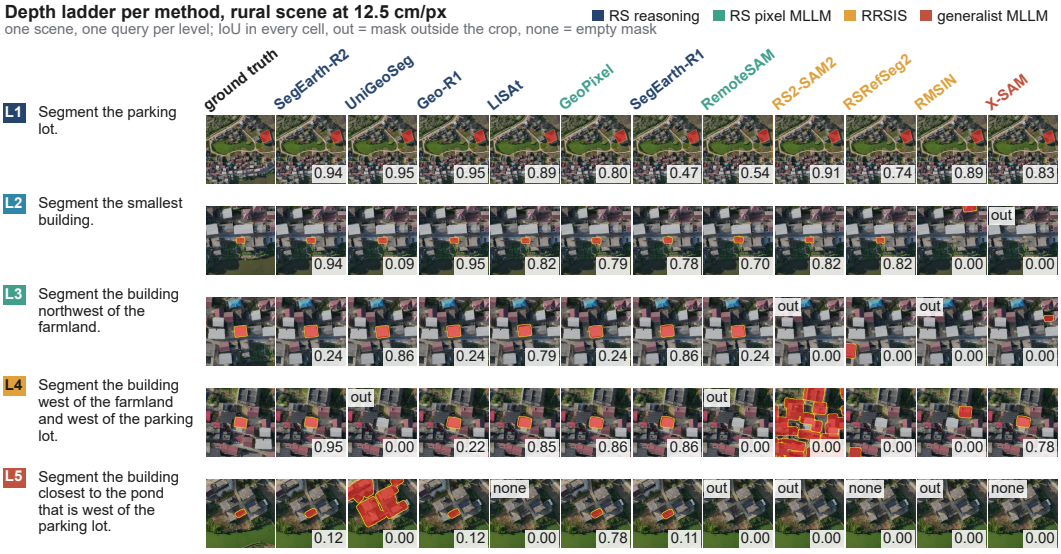


Figure 15: The depth ladder on a rural scene. Reading down the column shows whether the cost of depth is a property of the sentence or of the scene it was drawn from.

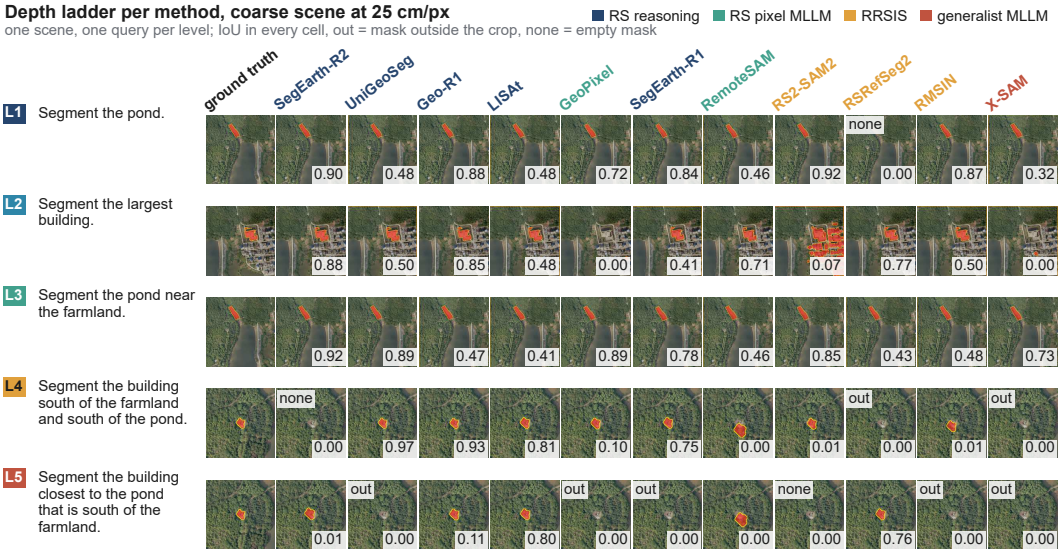


Figure 16: The depth ladder at 25 cm ground sampling, with the sentences held fixed, which separates language difficulty from pixel count.

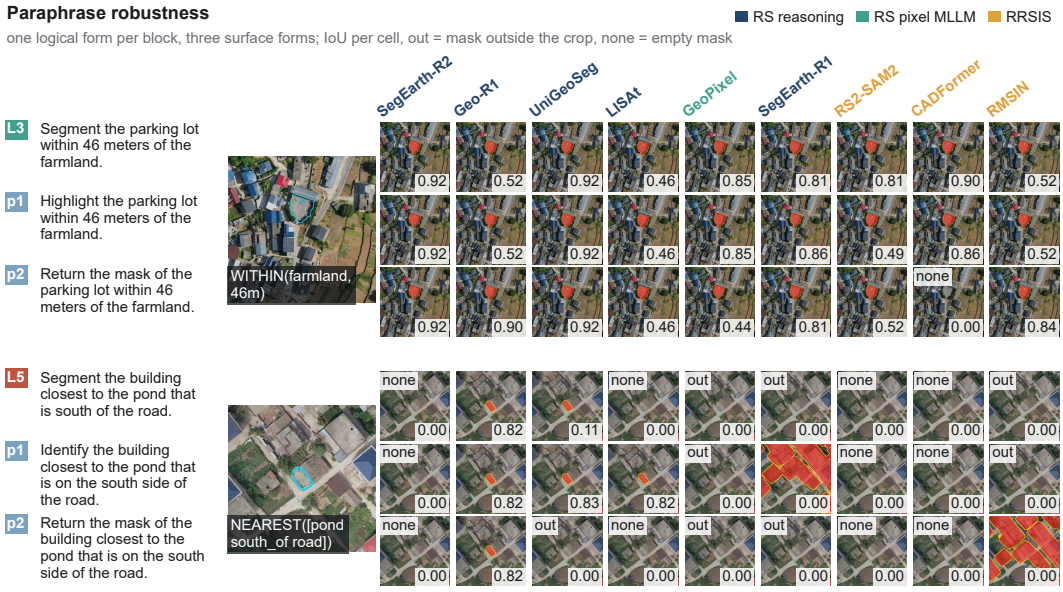


Figure 17: Paraphrase stability: one logical form, three surface forms, one row per method. A prediction that changes across wordings is a language failure and not a segmentation one, which is the measurement the three released wordings exist to support.

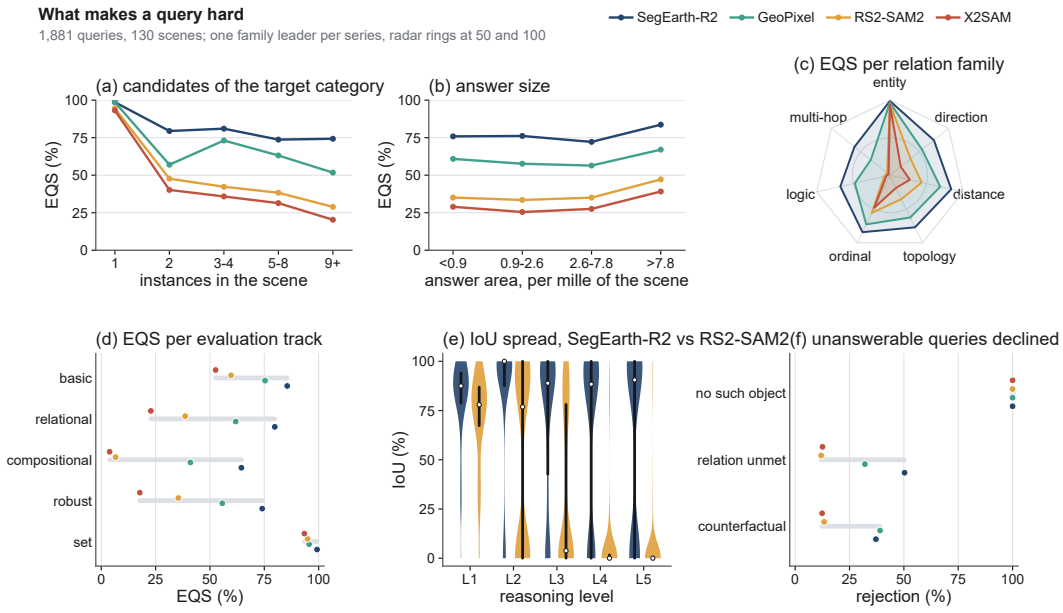


Figure 18: What makes a query hard: EQS against the number of same-category candidates, the answer's share of the scene, the relation family and the evaluation track, with the IoU spread and rejection behaviour beside them.

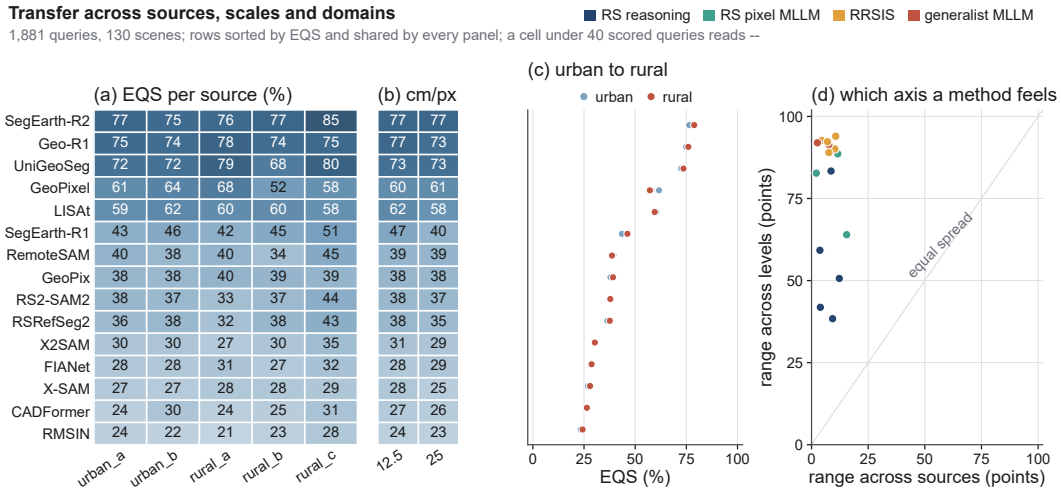


Figure 19: Generalisation across the axes the splits isolate: per source orthophoto, per ground sampling, urban to rural, and the spread a single mean hides.

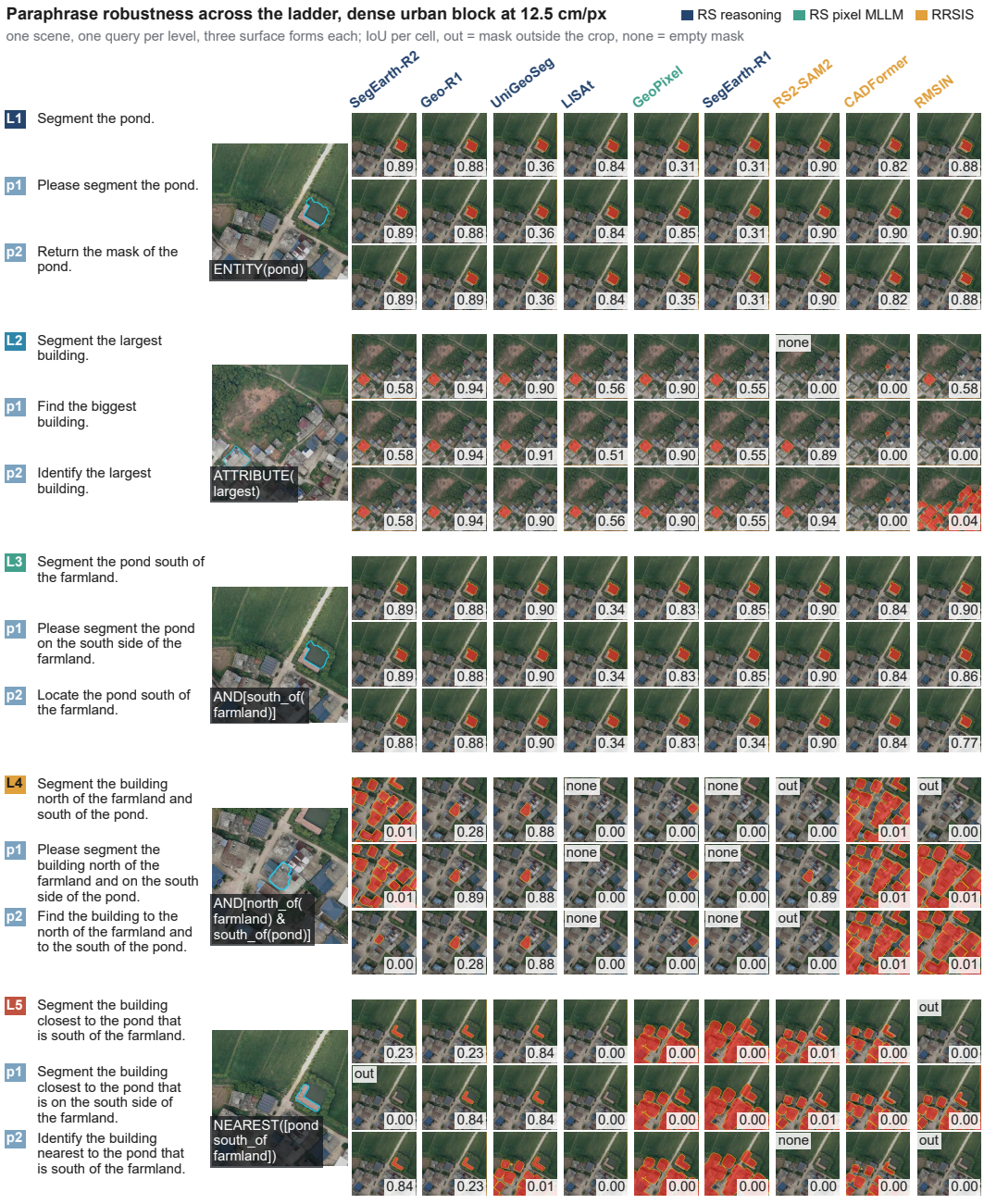


Figure 20: Paraphrase robustness across the ladder, dense urban block at 12.5 cm. One query per reasoning level, three surface forms each, IoU in every cell. Rows within a block share a logical form, printed in the scene panel, so a cell that changes down a block changes for the wording alone.




Figure 21: Paraphrase robustness across the ladder, farmland and ponds at 12.5 cm. The rural scenes are where a relation carries the most weight: the candidate sets are near-identical ponds and sheds, so the only thing separating the answer from its siblings is the clause a rewrite has to preserve.




Figure 22: Paraphrase robustness across the ladder, small-town blocks at 12.5 cm. Reading the L4 and L5 blocks against L1 is the point of the plate: a column may be identical across three wordings of a one-clause sentence and disagree with itself across three wordings of a two-relation one.

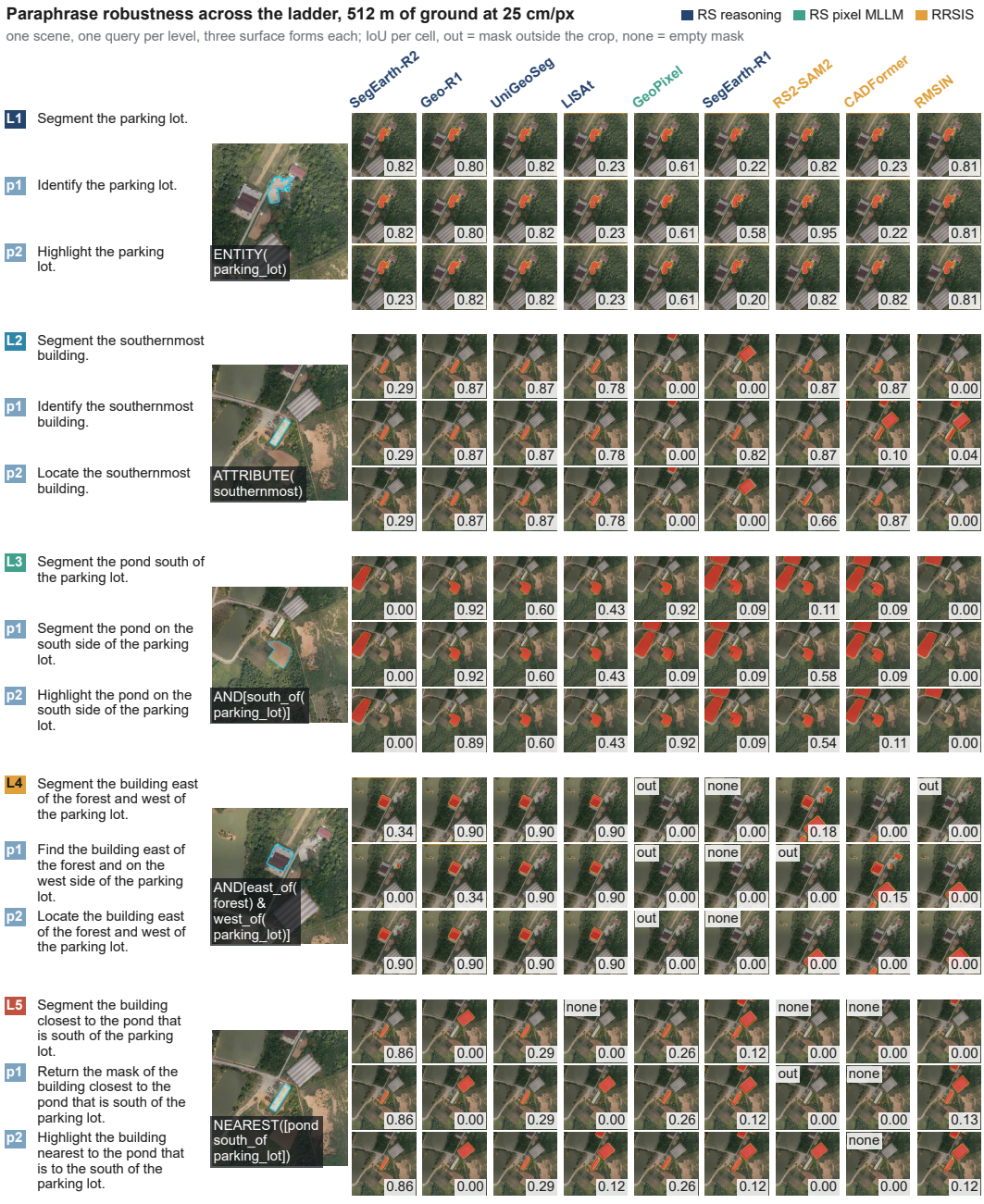


Figure 23: Paraphrase robustness across the ladder at 25 cm, over 512 m of ground. The coarse track halves every referent's pixel extent, so this plate separates a wording failure from a sampling one: a cell that is wrong at both ground samplings under every phrasing is not a language problem.