RIGHT, BUT CAN YOU SHOW IT? CORRECT ANSWERS FROM MULTIMODAL LLMS RARELY COME WITH FAITHFUL EVIDENCE

Anonymous authors
Paper under double-blind review

ABSTRACT

When a multimodal large language model (MLLM) answers a question about an image, it usually also states what it sees, such as which objects are present and where, as evidence for its answer. Users rely on this evidence to decide whether to trust the answer, yet benchmarks grade only the answer. They tacitly assume that *a correct answer comes with trustworthy evidence*. Testing this assumption means checking every stated fact against the image, and being able to prove a false fact false. That requires knowing everything the image contains: if a dataset labels only some of its objects, a claim about an unlabeled one might still be true and cannot be marked wrong, and most image datasets are labeled this way. Remote sensing is a rare exception. By protocol, its datasets label every object of declared categories, together with their spatial relations and their changes over time. Building on this, we introduce *evidence contracts*, which specify for each question the visual facts its answer depends on, and **RSFaith-Bench**, 13,511 human-validated questions over 16,288 images on which evidence is *checked against ground-truth annotation rather than judged by another model*. This check agrees with human raters on 89.4% of responses. Evaluating 22 MLLMs shows that the assumption fails in every one of them. Taken together, **more than half of the answers that accuracy counts as correct would fail a check of their own evidence.** Surprisingly, the dominant failure is not contradiction but missing support: most of these answers state nothing that the annotation refutes, yet their evidence does not establish the answer, a failure that no check of the stated claims alone can detect. More strikingly, **stronger models do not close the gap**. Across six larger or newer model versions, answer accuracy improves more than five times as much as faithful accuracy, and in half of them faithful accuracy *declines*. Human raters blind to our labels prefer the faithfully supported answer in up to 82% of matched pairs. Answer accuracy tells us when a model is right; it cannot tell us when to believe it. The benchmark and code are publicly available.

1 INTRODUCTION

When a multimodal large language model (MLLM) answers a question about an image, it rarely gives the answer alone. It also states what it sees: which objects are present, where they are, how they relate, and what has changed. This stated evidence matters most where the answer drives a decision. In remote sensing, MLLMs now support disaster-response triage, infrastructure change alerts, and environmental monitoring (Zhang et al., 2024; Kuckreja et al., 2024; Xuan et al., 2025; Wang et al., 2025c; Revankar et al., 2025), and their responses are read as an answer together with the visual facts behind it (Shabbir et al., 2025; Wang et al., 2026a; Liu et al., 2026; Li et al., 2026a). An analyst reviewing hundreds of such responses cannot re-inspect every image; the stated evidence is what tells her whether an answer can be trusted. Benchmarks, however, grade only the answer (Lobry et al., 2020; Li et al., 2024a;b; Wang et al., 2025b). In doing so, they rest on an assumption they never test: *a correct answer comes with correct evidence*.

A model can violate this assumption while still being scored correct. It may select the correct vessel type while asserting a spatial relation the image contradicts (Figure 1), or correctly report that buildings increased between two dates while stating only the later count, from which no increase follows.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

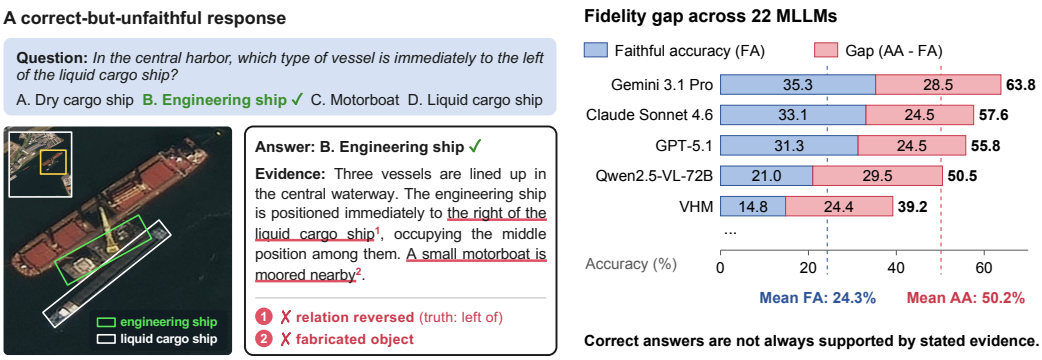


Figure 1: Correct-but-unfaithful response and fidelity gap. The model selects the correct vessel type, but its evidence reverses the spatial relation and adds unsupported content. Across 22 MLLMs, fewer than half of correct answers are faithfully supported.

We call such responses **correct-but-unfaithful**: *contradiction-bearing* when the evidence contains a refuted fact, and *missing-support* when the evidence does not establish the answer. Hallucination benchmarks, which check stated claims one at a time (Rohrbach et al., 2018; Li et al., 2023b; Jing et al., 2024; Wang et al., 2023), can flag the first but pass the second, whose error lies not in any claim it states but in what its claims fail to establish. Chain-of-thought faithfulness studies ask whether the stated reasoning caused the answer (Lanham et al., 2023; Turpin et al., 2023); we ask instead whether the stated evidence is true and sufficient, which is what a reader can actually check.

Why has this not been tested? Detecting missing evidence requires knowing which facts each answer depends on. Detecting false evidence is harder: it requires proving a claim false, and hence exhaustive annotation. If a dataset labels only some of its objects, as most image datasets do, a claimed object without a label may simply have been missed. Remote sensing is a rare exception. By protocol, object datasets annotate every instance of their declared categories (Ding et al., 2022; Sun et al., 2022), relation datasets use closed predicate inventories (Yan et al., 2025), and bi-temporal datasets record how each region changes (Yang et al., 2022; Shen et al., 2021). We turn this property into a verification procedure. For each question, an *evidence contract* specifies the visual facts its answer depends on, and each stated claim is checked against source annotation inside *closed scopes*, the predicates, regions, and time anchors within which the annotation is complete. Inside a closed scope a claim can be confirmed or refuted; outside it, the claim is marked uncertain. Evidence faithfulness thus becomes a quantity that is *computed rather than judged*.

On this basis we build **RSFaith-Bench**, 13,511 human-validated questions over 16,288 images from six public datasets, spanning perception, relational, and temporal reasoning, and evaluate 22 proprietary, open-source, and remote sensing MLLMs. A response counts toward *faithful accuracy* (FA) only if its answer is correct, none of its evidence is refuted, and its evidence fills the contract. The results overturn the assumption. Mean answer accuracy (AA) is 50.2% but FA only 24.3%: **more than half of correct answers fail a check of their own evidence**, and no model faithfully supports more than 61% of its correct answers. Surprisingly, **the dominant failure is missing support rather than contradiction**: missing-support responses account for 63% of the gap, and no check of the stated claims alone can catch them. More strikingly, **stronger models do not close the gap**: across six larger or newer versions within a model family, AA rises by 3.9 points on average while FA rises by only 0.7, and declines in half of them. Ranking by FA reorders models, and human raters prefer faithfully supported answers over equally correct unfaithful ones in up to 82% of pairs.

Answer accuracy tells us when a model is right. It does not tell us when to believe it; for that, the evidence itself must be measured. Our contributions are threefold:

- **A verifiable definition of answer–evidence faithfulness.** Evidence contracts and a three-valued verifier over closed scopes turn faithfulness from a quantity judged by humans or models into one computed from annotation, in any domain whose annotation declares where it is complete.
- **RSFaith-Bench, a rigorously validated benchmark.** 13,511 human-validated questions (inter-annotator agreement: $\kappa = 0.81$), with an automated verdict that agrees with schema-blind human judgment on 89.4% of responses and remains stable across extractors and prompt formats.

- **Evidence that being right and showing it are distinct capabilities.** A 22-model study that exposes a gap invisible to both answer accuracy and hallucination checks, persistent across model scale, and consequential for model selection and human trust.

2 Related Work

2.1 Faithfulness and Hallucination Evaluation

Work on whether multimodal outputs are faithful to the image differs in what it checks. Object hallucination benchmarks test whether described objects are present (Rohrbach et al., 2018; Li et al., 2023b), including for remote sensing imagery (Zhou et al., 2026), and claim-level methods decompose responses into atomic statements and verify each against the image (Jing et al., 2024; Wang et al., 2023; Yin et al., 2026; Cao et al., 2025). Some of these checks ask a vision-language model to judge each claim (Jing et al., 2024; Zhou et al., 2026), and those that rely on annotation can refute at most object presence or differences introduced by image editing (Table A1). A second line evaluates the reasoning around an answer: the consistency of intermediate traces (Jiang et al., 2025; Dong et al., 2025), whether a multimodal chain of thought agrees with the visual evidence it cites (Lv et al., 2026), and whether cited evidence is anchored on relevant image regions (Wang et al., 2026b; Li et al., 2026b), where correct answers are observed to rest on irrelevant regions (Li et al., 2026b). Chain-of-thought faithfulness studies ask whether stated reasoning reflects the computation that produced the answer, typically by intervening on the reasoning (Lanham et al., 2023; Turpin et al., 2023; Tanneru et al., 2024). Answer-conditioned sufficiency, whether a rationale contains what the answer requires, has been formalized for text rationales (DeYoung et al., 2020), where it is measured through the model's own predictions. These directions score a claim, a trace, or a region. None checks, against annotation that can refute a claim, whether the stated visual facts suffice for the answer they accompany: a response can contain only true claims yet omit the facts the answer depends on, and a correct answer can come with a refuted fact that claim-level scoring never ties to the answer. RSFaith-Bench makes this answer–evidence relation the unit of evaluation. Where grounded-reasoning benchmarks ask where a model looks, it asks whether what the model states establishes the answer.

2.2 Remote Sensing Benchmarks

Evaluation of vision-language models on remote sensing imagery has progressed from visual question answering over aerial images (Lobry et al., 2020; Li et al., 2024a) to benchmarks covering scene understanding, ultra-high-resolution imagery, and reference grounding (Li et al., 2024b; Danish et al., 2025; Wang et al., 2025b), multi-temporal reasoning (Xuan et al., 2025; Luo et al., 2026), disaster damage assessment (Wang et al., 2025c), spatial and 3D reasoning (Tsujimoto et al., 2026; Xu et al., 2026), and pixel-grounded Earth observation (Fu et al., 2026; Shu et al., 2026). These benchmarks score the final answer, a grounding region, or a pixel-level prediction (Table A1). Their source annotations, however, are complete by protocol within declared categories, predicates, and time anchors (Section 1), which lets a stated claim be refuted as well as confirmed. RSFaith-Bench builds its closed scopes on this property (Section 3.1).

3 RSFaith-Bench

Each RSFaith-Bench item pairs an image or a bi-temporal image pair with a multiple-choice question and its gold answer, and the model states the relevant visual evidence before answering. A response is faithful when the answer is correct, no stated visual claim conflicts with the source annotation, and the supported claims include every fact the answer depends on. Figure 2 traces how these conditions are checked: a grounded fact layer labels each stated claim (Section 3.1), an evidence contract specifies the facts each answer depends on (Section 3.2), and the scoring protocol combines the two (Section 3.3).

We use *faithfulness* in the sense of consistency with the source, as in summarization (Maynez et al., 2020) and claim-level visual faithfulness (Jing et al., 2024), extended to answer sufficiency: the

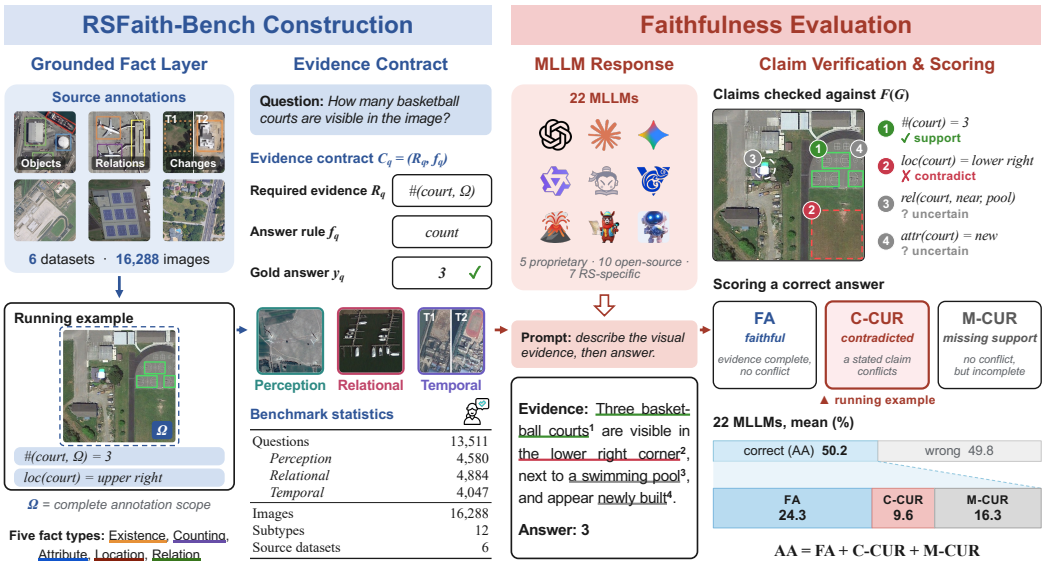


Figure 2: Overview of RSFaith-Bench. Left: source annotations form a grounded fact layer $F(G)$, and each question carries an evidence contract. Right: claims in the response are verified against $F(G)$ and matched to the contract; the answer shown is correct, but one refuted claim makes the response contradiction-bearing (C-CUR).

stated evidence must agree with the source annotation wherever that annotation is complete, and the claims it confirms must establish the answer. This makes faithfulness a property of the stated evidence, complementary to chain-of-thought faithfulness, which studies the reasoning process (Jacovi & Goldberg, 2020; Lanham et al., 2023).

3.1 Grounded Fact Layer

Data sources. All checks are made against source annotation, so every verdict traces back to an annotated record. RSFaith-Bench uses the held-out evaluation splits of six public remote sensing datasets spanning the three reasoning levels: DIOR (Li et al., 2020), DOTA-v2.0 (Ding et al., 2022), and FAIR1M (Sun et al., 2022) provide oriented object instances with fine-grained categories (Perception); ReCon1M (Yan et al., 2025) provides relation triplets between objects (Relational Reasoning); and SECOND (Yang et al., 2022) and xBD (Gupta et al., 2019) provide bi-temporal land-cover transitions and building damage states (Temporal Reasoning; Appendix A.1).

Fact layer. The six datasets use different annotation formats, so we convert each image or image pair into a grounded scene graph $G = (V, E, A)$ whose vertices, edges, and attributes keep links to their source records. The graph yields a fact layer $F(G)$ of typed facts over five predicates: Existence, Counting, Attribute, Location, and Relation. Gold answers, evidence contracts, and model claims are all expressed with these predicates, so they can be compared directly. In Figure 2, $F(G)$ records three basketball courts in the upper right of the image.

Closed scopes. A matching fact confirms a claim, but refuting a claim requires knowing that the annotation is complete where the claim applies. We record this completeness as closed scopes: a closed scope $\Omega = (p, \sigma, a, t)$ names a predicate, its argument types, a spatial region, and a time anchor for which the source annotation is complete, such as the declared categories of an object dataset, the declared relation predicates of ReCon1M, or the labeled states of a bi-temporal pair. Verification then follows a three-valued rule: a claim is labeled `support` if it matches a fact in $F(G)$, `contradict` if a closed scope records an incompatible fact or a different exact count, and `uncertain` otherwise. In Figure 2, both the count and the location of the courts are covered by closed scopes, so a claim about either can be confirmed or refuted. Every contradiction is thus backed by a record in the same annotation used to build the benchmark, and the source links in G point to it (Appendix A.3).

3.2 Evidence Contracts and Question Generation

Evidence contracts. The fact layer decides whether a claim is supported, but judging whether the supported claims suffice for an answer requires an exact statement of the facts the answer depends on. We write this statement first and generate the question from it, so the gold answer and its evidence requirements come from the same facts. Each question q carries an evidence contract $C_q = (R_q, f_q)$, where R_q is a set of typed evidence slots, each binding a predicate to a verifiable scope in $F(G)$, and f_q maps a complete slot assignment to an answer; the gold assignment R_q^* is read from $F(G)$, and $y_q = f_q(R_q^*)$. Facts fixed by the question text, such as the queried region or time anchor, need not be restated unless they are themselves queried. For the counting question in Figure 2, R_q holds one COUNTING slot $\#(\text{court}, \Omega)$, f_q returns its value, and $y_q = 3$.

Reasoning levels. The three reasoning levels differ in the evidence a faithful answer must state (Figure 3; Figure 8 shows one example per subtype).

Perception covers questions about objects in a single image: object presence, counting, localization, and fine-grained recognition (Johnson et al., 2017; Hudson & Manning, 2019). The counting question of Figure 2 belongs to this level, and the only evidence it requires is the number of basketball courts in the image.

Relational Reasoning tests how objects or regions are arranged. Directional, topological (such as containment), proximity, and projective-ordering questions follow the standard families of qualitative spatial relations (Cohn & Renz, 2008; Hernández, 1994), and aggregate-distribution questions ask where a category is concentrated. Every object that a question mentions is uniquely identifiable in the image, and a faithful answer to "Which listed object is to the right of the building?" must identify both objects and state where the chosen object lies relative to the building.

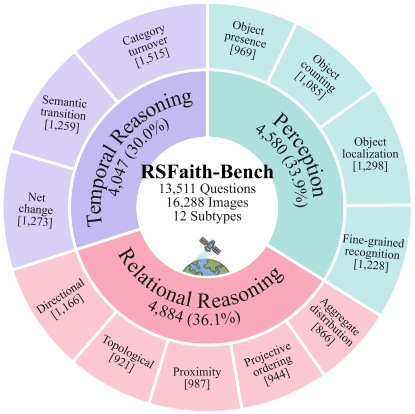


Figure 3: RSFaith-Bench composition by reasoning level and question subtype. Brackets show question counts; percentages show level shares.

Temporal Reasoning questions compare two images of the same area taken at different dates: which category appears or disappears (category turnover), how the category of a region or building changes (semantic transition), or which category grows or shrinks the most (net change) (Singh, 1989; Yang et al., 2022). The evidence must therefore cover both dates; reporting only the later state leaves the contract unfilled even when the answer is correct.

Question programs. An executable program over $F(G)$ generates each question together with its answer options, gold answer, and contract. A question is generated only when every slot in R_q can be filled from closed scopes and its answer is unique. Counting and aggregate questions use bounded count ranges, and ambiguous references, borderline geometric relations, and unresolved temporal correspondences are removed. Every question is thus decidable from the annotation, both its answer and each fact its contract requires (Appendix A.2).

Human validation. Each of the 13,511 questions is validated by an annotator with graduate-level training in remote sensing or computer vision, who checks the question, gold answer, and evidence contract together against the source image and annotation, correcting local errors and removing unresolved cases. After calibration on a stratified pilot, the three annotators reach $\kappa = 0.81$ on a multiply annotated held-out set, and validation totals approximately 337 annotator-hours (Appendix A.4).

3.3 Faithfulness Evaluation

Scoring follows the right half of Figure 2: the visual claims in a response are extracted, verified against $F(G)$, and matched against the contract.

Claim extraction. A response is split into the stated answer $\hat{y}_q$ and the evidence text $\hat{t}_q$. The answer is mapped into the answer space of f_q, and answers that cannot be mapped count as incorrect. A single extractor, frozen before evaluation, reads only the evidence text, without the image or the

Table 1: Main results on RSFaith-Bench (%). CC and MC abbreviate C-CUR and M-CUR, so AA = FA + CC + MC; CP is claim precision. **Bold**/underlined: best/second-best AA and FA; **red bold**/red underlined: largest/second-largest CC and MC.

Model	Overall					Perception					Relational					Temporal				
	AA↑	FA↑	CC↓	MC↓	CP↑	AA	FA	CC	MC	CP	AA	FA	CC	MC	CP	AA	FA	CC	MC	CP
Proprietary MLLMs																				
GPT-4o	54.5	30.0	5.5	19.0	80.5	61.0	33.4	6.5	21.1	80.4	53.8	32.2	2.0	19.5	82.7	48.1	23.6	8.5	15.9	77.2
GPT-5.1	55.8	31.3	11.8	12.7	77.2	63.0	33.8	13.4	15.9	74.4	55.9	32.8	8.4	14.7	80.8	47.6	26.7	14.1	6.8	74.1
Claude Haiku 4.5	46.4	28.3	11.0	7.1	71.5	49.1	28.4	12.8	7.8	68.2	48.5	33.8	6.2	8.5	75.4	40.9	21.5	14.7	4.7	68.4
Claude Sonnet 4.6	57.6	33.1	12.6	11.8	73.6	61.4	36.7	13.7	11.0	69.9	60.0	**40.3**	5.9	13.8	79.3	50.3	20.4	**19.6**	10.3	70.1
Gemini 3.1 Pro	**63.8**	**35.3**	10.7	17.8	80.1	**70.0**	35.2	13.8	21.1	76.1	**65.4**	35.5	6.1	23.9	83.6	**54.8**	**35.3**	12.8	6.7	78.3
Open-source MLLMs																				
Qwen3-VL-8B	52.4	27.3	13.2	12.0	74.8	60.1	31.3	16.2	12.6	73.8	52.9	30.5	9.2	13.2	77.8	43.2	18.8	14.6	9.9	71.3
Qwen3-VL-32B	55.5	29.9	**17.4**	8.3	73.9	61.6	33.3	**18.8**	9.6	71.1	56.4	33.1	**14.7**	8.6	78.9	47.7	22.1	19.0	6.6	68.5
Qwen3.5-9B	53.4	28.4	10.0	15.1	76.8	58.2	36.1	10.0	12.2	75.2	55.4	29.4	7.7	18.4	80.7	45.6	18.5	12.8	14.3	72.9
Qwen3.5-27B	57.8	27.1	13.7	17.0	76.7	62.5	29.7	16.5	16.2	74.8	60.2	28.0	10.4	21.7	81.1	49.8	23.1	14.5	12.2	72.4
Qwen2.5-VL-7B	47.6	22.4	8.9	16.3	76.1	55.6	26.3	7.8	21.4	72.6	47.8	23.9	9.6	14.3	78.4	38.3	16.1	9.2	13.0	74.1
Qwen2.5-VL-72B	50.5	21.0	13.7	15.8	74.8	55.8	23.2	16.3	16.3	73.0	54.8	21.3	14.0	19.5	77.8	39.1	18.0	10.3	10.8	71.8
InternVL3.5-8B	55.2	31.9	6.1	17.2	78.9	67.1	41.1	10.1	15.9	80.6	57.3	33.5	2.0	21.8	81.3	39.4	19.6	6.5	13.3	74.3
InternVL3.5-38B	55.9	30.3	6.9	18.6	79.2	67.1	41.8	8.9	16.4	80.7	58.3	27.4	4.1	26.8	83.0	40.4	21.0	8.1	11.3	72.3
LLaVA-OneVision-7B	47.6	12.7	8.6	26.2	79.9	54.7	22.6	11.8	20.3	78.8	47.6	8.4	5.5	33.7	83.5	39.5	6.7	8.7	**24.0**	75.9
GLM-4.5V	56.2	33.8	10.0	12.5	78.0	62.7	36.1	9.4	17.2	76.2	55.5	37.6	5.6	12.3	81.8	49.8	26.5	15.9	7.5	74.0
Remote sensing MLLMs																				
RSThinker	42.0	24.2	6.7	11.1	78.1	56.2	36.8	8.6	10.8	77.4	38.1	24.8	2.9	10.5	80.1	30.5	9.1	9.1	12.3	75.3
GeoChat	35.9	9.2	4.5	22.1	71.5	40.4	24.3	5.5	10.6	74.4	36.4	2.0	3.1	31.3	79.6	30.2	0.9	5.2	**24.0**	60.4
VHM	39.2	14.8	4.8	19.6	76.7	54.1	33.3	10.4	10.4	72.4	34.2	5.0	0.6	28.6	79.8	28.4	5.7	3.6	19.1	76.8
GeoReason	49.0	17.8	10.0	21.2	73.2	55.6	30.7	8.6	16.3	74.0	47.1	7.4	8.9	30.7	72.9	43.9	15.8	13.0	15.1	73.0
GeoZero-8B	49.3	22.2	8.5	18.6	67.1	56.6	**45.3**	5.2	6.0	71.6	46.2	11.1	2.6	32.4	67.4	44.9	9.2	19.4	16.2	63.0
RemoteReasoner	47.9	14.4	6.8	**26.7**	77.7	53.1	21.3	7.8	**23.9**	78.1	52.6	11.9	5.8	**34.8**	77.9	36.3	9.5	6.8	20.0	76.8
RS-EoT-7B	30.1	8.7	9.8	11.6	63.7	43.5	18.5	15.0	9.9	61.5	32.3	2.7	8.4	21.3	67.7	12.3	5.0	5.4	1.9	60.6

contract. It separates visual perception from analytical text, splits the visual part into atomic claims, and normalizes each claim to the five predicates of $F(G)$; analytical content, such as the general statement that airports usually hold many aircraft, is not verified (Appendix B.2).

Verification. Each normalized claim receives the three-valued label of Section 3.1, which yields the sets E_q^+ (supported), E_q^- (contradicted), and E_q^u (uncertain). Across the evaluated models, a median of 68% of extracted claims (interquartile range 62–72%) receive a `support` or `contradict` label; the remainder are `uncertain` and do not count as contradictions (Appendix D.7).

Slot filling. Supported claims are matched against the contract: a claim fills a slot of R_q when its type, scope, predicate, and value match the slot, a counting slot is also filled when supported counts over disjoint subregions covering its scope sum to its value, and $\mathrm{Fill}(R_q, E_q^+)$ is the set of complete slot assignments obtained this way. In Figure 2, "three basketball courts" fills the counting slot, "in the lower right corner" is contradicted by the recorded location, and the annotation cannot decide "next to a swimming pool" or "newly built", which remain uncertain.

Metrics. *Answer Accuracy* (AA) is the fraction of questions answered correctly, $\hat{y}_q = y_q$. *Faithful Accuracy* (FA) is the fraction of questions for which (i) the answer is correct, (ii) the response contains no contradicted claim, and (iii) some slot assignment built from the supported claims derives the gold answer:

$$\mathrm{FA} = |Q|^{-1} \sum_q \mathbf{1}\big[\, \hat{y}_q = y_q \;\wedge\; |E_q^-| = 0 \;\wedge\; \exists A \in \mathrm{Fill}(R_q, E_q^+) \colon f_q(A) = y_q \,\big]. \tag{1}$$

The *fidelity gap* is $\mathrm{AA} - \mathrm{FA} \geq 0$, and the *faithful share* is FA/AA. The gap splits into two rates: *C-CUR* (contradiction-bearing correct-but-unfaithful), the share of questions answered correctly with at least one contradicted claim ($|E_q^-| > 0$), and *M-CUR* (missing-support correct-but-unfaithful), the share answered correctly with no contradicted claim but without an assignment that derives the gold answer, so that $\mathrm{AA} = \mathrm{FA} + \text{C-CUR} + \text{M-CUR}$. *Claim Precision* (CP) is the share of supported claims among all claims the verifier decides, pooled over questions: $\mathrm{CP} = \sum_q |E_q^+| \,/\, \sum_q (|E_q^+| + |E_q^-|)$.

4 Experiments

4.1 Experimental Setup

We evaluate 22 MLLMs, grouped as in Table 1:

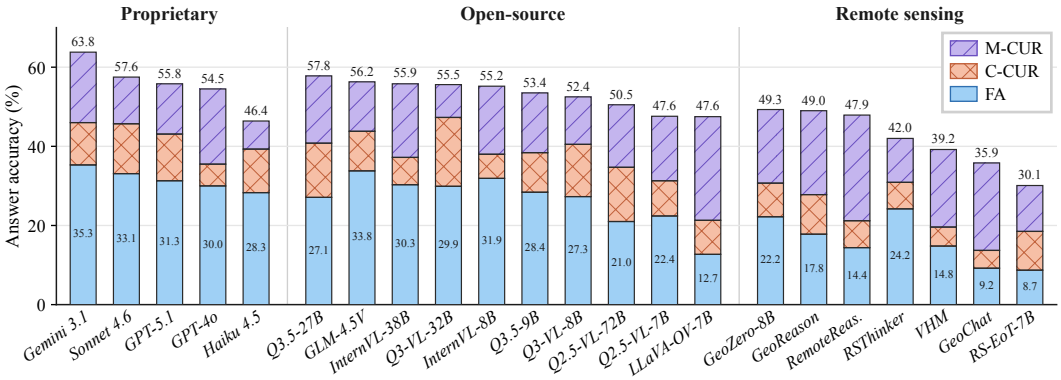


Figure 4: Answer accuracy decomposed by evidence status. AA is partitioned into faithfully supported correct responses (FA), contradiction-bearing correct responses (C-CUR), and missing-support correct responses (M-CUR).

- **Proprietary MLLMs**: GPT-4o (OpenAI, 2024), GPT-5.1 (OpenAI, 2025), Claude Haiku 4.5 (Anthropic, 2025), Claude Sonnet 4.6 (Anthropic, 2026), and Gemini 3.1 Pro (Google DeepMind, 2026).
- **Open-source MLLMs**: Qwen3-VL-8B/32B (Bai et al., 2025a), Qwen3.5-9B/27B (Qwen Team, 2026), Qwen2.5-VL-7B/72B (Bai et al., 2025b), InternVL3.5-8B/38B (Wang et al., 2025d), LLaVA-OneVision-7B (Li et al., 2025), and GLM-4.5V (GLM-V Team et al., 2025).
- **Remote sensing MLLMs**: RSThinker (Liu et al., 2026), GeoChat (Kuckreja et al., 2024), and VHM (Pang et al., 2025), together with four recent reasoning-oriented models: GeoReason (Li et al., 2026a), GeoZero-8B (Wang et al., 2025a), RemoteReasoner (Yao et al., 2026), and RS-EoT-7B (Shao et al., 2026).

Evaluation protocol. All models are evaluated zero-shot under identical conditions: the same prompt template, which asks for the visual evidence before the answer (Appendix B.2), and each model's default generation configuration (Appendix B.1). Qwen3-32B (Yang et al., 2025) serves as the frozen claim extractor, and the verifier labels each extracted claim as `support`, `contradict`, or `uncertain`. We report the metrics of Section 3.3 overall and for each reasoning level; Section 4.4 reports reliability checks for this pipeline.

4.2 Main Results

More than half of correct answers fail a check of their own evidence. Table 1 reports overall and level-wise results. Mean AA is 50.2% and mean FA 24.3%, so fewer than half of the answers that answer-only evaluation scores as correct would pass a check of their own stated evidence. Every model has a gap, from 17.8 (RSThinker) to 34.9 points (LLaVA-OneVision-7B), and the gap holds at every reasoning level (25.9 points for Perception, 27.5 for Relational, and 24.0 for Temporal Reasoning) and within every model class. No model's faithful share exceeds 61.0% (Claude Haiku 4.5), and even Gemini 3.1 Pro, which has the highest AA and the highest FA, leaves 45% of its correct answers unfaithful. FA is attainable: the best models per subtype reach 78% on object localization and 70% on object presence (Appendix C.1).

Table 2: Upgrades within a model family to a larger or newer version (%).

Upgrade	ΔAA	ΔFA	ΔGap
Qwen2.5-VL 7B → 72B	+2.9	−1.4	+4.3
Qwen3-VL 8B → 32B	+3.1	+2.6	+0.5
Qwen3.5 9B → 27B	+4.4	−1.3	+5.7
InternVL3.5 8B → 38B	+0.7	−1.6	+2.3
GPT-4o → GPT-5.1	+1.3	+1.3	0.0
Claude Haiku 4.5 → Sonnet 4.6	+11.2	+4.8	+6.4
Mean	+3.9	+0.7	+3.2

Missing support, not contradiction, accounts for most of the gap. Figure 4 splits each model's AA into FA, C-CUR, and M-CUR. M-CUR carries 63% of the mean gap and exceeds C-CUR in 18 of 22 models; the four exceptions, both Claude and both Qwen3-VL models, are among the most verbose (Section 4.3). Eleven models have C-CUR of at least 10%, spanning all three model classes. A C-CUR response contains a claim refuted by the verifier, whereas an M-CUR response contains no such claim but lacks enough supported evidence to fill the contract. Contract coverage thus identifies a failure that contradiction checks alone miss, and the contract-aware human audit,

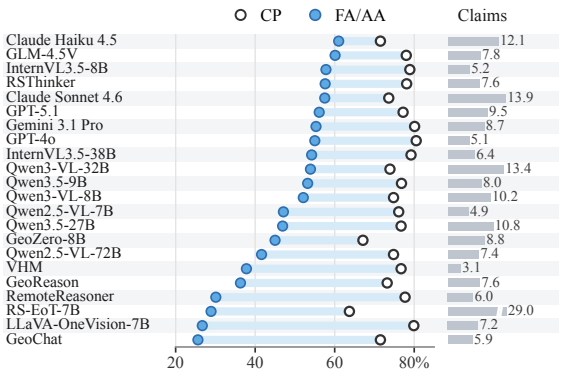


Figure 5: Claim precision CP (○) and faithful share FA/AA (●) per model, sorted by FA/AA; grey bars give the average number of extracted claims per response.

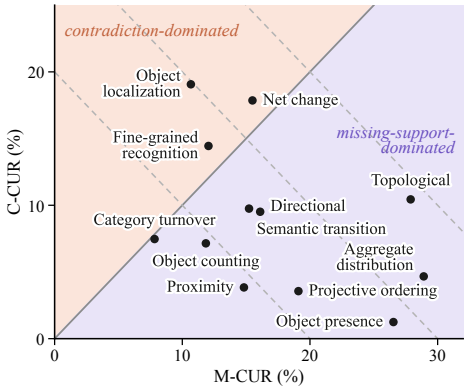


Figure 6: Subtype failure map (22-model mean); above the diagonal C-CUR exceeds M-CUR, and dashed lines mark 20-, 30-, and 40-point gaps.

which redoes extraction, verification, and slot filling, changes M-CUR by at most 0.3 points per model (Table 3).

Stronger models do not close the gap. Table 2 follows six upgrades within a model family, each to a larger or newer version. AA rises in all six, by 3.9 points on average, while FA rises by only 0.7 and falls in three of them; the gap widens in five upgrades and is unchanged in the sixth. For the four open-source families, where only the model size changes, AA rises by 2.8 points on average and FA falls by 0.4. Progress on answers does not transfer to evidence.

Ranking by FA reorders models. Eight of the 22 models move by at least three positions when sorted by FA rather than AA. The largest individual reorderings include Qwen3.5-27B, which ranks second on AA (57.8%) but falls outside the top ten on FA (27.1%), Claude Haiku 4.5, which ranks eighteenth on AA (46.4%) and tenth on FA (28.3%), and RSThinker, which ranks nineteenth on AA (42.0%) and thirteenth on FA (24.2%). The pattern also appears within reasoning levels: in Relational Reasoning, Claude Sonnet 4.6 reaches FA = 40.3% with AA = 60.0%, while Gemini 3.1 Pro reaches FA = 35.5% with AA = 65.4%.

4.3 Fine-Grained Analysis

FA carries a signal that claim precision does not. Claim precision (CP) lies between 71.5% and 80.5% for 20 of the 22 models (GeoZero-8B: 67.1%, RS-EoT-7B: 63.7%), while FA/AA ranges from 25.6% to 61.0% (Figure 5). GPT-4o and Gemini 3.1 Pro have nearly identical CP (80.5 versus 80.1) but differ by 5.3 FA points. Stating true claims and stating the claims the answer depends on are different achievements, and FA measures the second.

Response length shapes how an unfaithful answer fails. We define CTR = C-CUR/(C-CUR + M-CUR), the share of unfaithful correct responses that contain at least one contradicted claim. CTR ranges from 0.17 to 0.68 and rises with the number of claims a model states (Spearman $\rho = 0.83$ across the 22 models): models that state more are more often caught asserting a refuted fact, and models that state less more often leave the contract unfilled. Qwen3-VL-32B (13.4 claims per response) and InternVL3.5-38B (6.4) reach similar AA (55.5% and 55.9%) with CTR of 0.68 and 0.27. FA rewards neither strategy. The faithful share FA/AA correlates with claims per response at only $\rho = 0.22$, and under a fixed budget of exactly five atomic visual claims on 500 matched questions, FA rankings retain Spearman $\rho = 0.87$ and four of the top five models, with a mean absolute FA change of 2.8 points (Appendix D.9). FA thus measures the quality of the stated evidence.

Answer options raise AA far more than FA. When the options are removed and models answer in their own words, mean AA across eight models on 500 matched questions falls from 52.4% to 38.4%, but mean FA only from 24.2% to 20.7% (Appendix D.10). Thus 75% of the drop in AA comes from correct answers that lacked faithful support; the options mostly add answers of this kind. The gap persists for open-ended questions, where every model retains 12.1–22.0 points and 46.2% of correct answers remain contradiction-bearing or missing-support.

Contradictions concentrate where the answer demands a detailed value. Across the 12 subtypes (Figure 6), average C-CUR ranges from 1.2% (object presence) to 19.1% (localization). The three subtypes with the highest average C-CUR are localization (19.1), net change (17.9), and fine-grained recognition (14.4). Each asks the model to commit to a detailed value: a spatial position, a count change between time anchors, or a fine-grained subtype label. The three subtypes with the lowest average C-CUR are object presence (1.2), projective ordering (3.6), and proximity (3.8), where the fidelity gaps are dominated by missing support. Object counting (average C-CUR 7.1) and net change (17.9) sharpen the contrast: both involve counts, but net change requires a cross-time count delta and produces a much larger contradiction component. Appendix E gives qualitative cases of faithful and contradiction-bearing evidence.

4.4 Reliability Checks

Human judgment. On 500 stratified questions (five models, 2,500 responses), three annotators independent of benchmark construction rate each response as fully, partially, or not supported without seeing the evidence slots or automated labels, two independently and the third adjudicating disagreements; the adjudicated, binarized ratings agree with FA on 89.4% of responses ($\kappa = 0.74$). A contract-aware audit of the same responses changes FA, C-CUR, and M-CUR by at most 0.8 points per model (Table 3; Appendix D.1). In 400 answer-matched pairs with similar claim counts and evidence lengths, raters blinded to model identity and labels prefer the faithful response in 81.5% of pairs against C-CUR and 69.0% against M-CUR (Appendix D.3).

Table 3: Contract-aware human audit of the automated scores (%); $\Delta = \text{Human} - \text{Auto}$.

Model	FA			C-CUR			M-CUR		
	Human	Auto	Δ	Human	Auto	Δ	Human	Auto	Δ
GPT-5.1	27.8	27.0	+0.8	16.7	17.2	−0.5	12.9	13.2	−0.3
Claude Sonnet 4.6	29.2	28.6	+0.6	14.7	15.0	−0.3	11.1	11.4	−0.3
Qwen2.5-VL-72B	20.7	20.2	+0.5	14.8	15.2	−0.4	14.7	14.8	−0.1
RSThinker	23.1	22.4	+0.7	6.8	7.4	−0.6	12.1	12.2	−0.1
GeoChat	9.0	8.6	+0.4	3.6	3.8	−0.2	21.0	21.2	−0.2

Visual dependence. On the same 500 questions, removing the image while keeping the question and prompt fixed lowers AA from 33.6–57.4% to 20.2–27.3% across six models (Figure 7), at or below the 25–33% chance level of the answer formats, so answering requires the image (Appendix D.4).

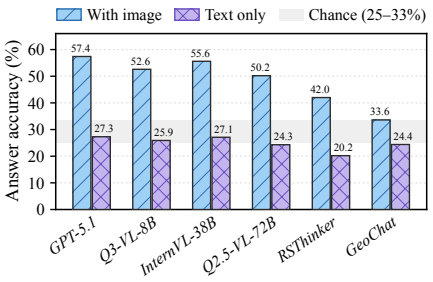


Figure 7: Answer accuracy of six models with and without the input image on the 500-question subset.

Pipeline robustness. The extractor reaches $F_1 = 0.91$ against independent gold claims for 500 responses, and auditors confirm 95.2% of sampled contradiction labels (Appendices D.5 and D.2). Replacing Qwen3-32B with GPT-4.1 on fixed responses to 500 questions from all 22 models changes FA by 1.20 points on average, with an FA rank correlation of 0.97 (Appendix D.6). Answer-first and open-format prompts change FA by at most 2.4 points for six models and leave their FA ordering unchanged (Appendix D.8).

5 Conclusion

This work introduced evidence contracts, which name the visual facts each answer depends on, and RSFaith-Bench, which checks the evidence a multimodal model states against ground-truth remote sensing annotation. Across 22 MLLMs, more than half of correct answers fail this check. Most failures are missing support: the stated facts do not establish the answer, a failure that checking claims one at a time cannot reveal. The rest state a fact that the annotation contradicts. Model upgrades raise answer accuracy while faithful accuracy stays nearly flat. Being right and showing why are distinct capabilities, and progress on the first does not bring the second. The test needs only annotation that declares where it is complete, so it extends to any domain with such annotation. As multimodal models enter decisions that people act on, the evidence they state deserves the same scrutiny as the answers they give, and evidence contracts make this scrutiny possible.

AI USE STATEMENT

Generative AI tools were used to assist with literature retrieval and the drafting and polishing of manuscript text. Within the evaluation method, Qwen3-32B serves as the frozen claim extractor and GPT-4.1 as an alternative extractor in the robustness analysis (Section 3.3, Appendix D.6). The authors reviewed all AI-assisted content and take full responsibility for the manuscript, including its text, claims, and artifacts.

ETHICS STATEMENT

All authors adhere to the ICLR Code of Ethics. RSFaith-Bench is built entirely from six public remote sensing datasets and involves no personal data. All human annotation was carried out voluntarily by members of the authors' research group with graduate-level training in remote sensing or computer vision, and the release follows the licenses and attribution requirements of the source datasets.

REPRODUCIBILITY STATEMENT

The construction pipeline is specified in Sections 3.1–3.2 and Appendix A, and Appendix B lists every model checkpoint or API snapshot, the verbatim prompts, the verification rules, and the metric computation. The claim extractor is an open-weight model frozen before evaluation and the verifier is rule-based, so all scores are reproducible without a proprietary judge. The benchmark and code are publicly available.

REFERENCES

Anthropic. Claude Haiku 4.5 system card. System card, 2025. URL https://www.anthropic.com/claude-haiku-4-5-system-card.

Anthropic. Claude Sonnet 4.6 system card. System card, 2026. URL https://www.anthropic.com/claude-sonnet-4-6-system-card.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, et al. Qwen3-VL technical report. Technical report, arXiv, 2025a. URL https://arxiv.org/abs/2511.21631.

Shuai Bai, Keqin Chen, Xuejing Liu, et al. Qwen2.5-VL technical report. Technical report, arXiv, 2025b. URL https://arxiv.org/abs/2502.13923.

Rui Cao, Zifeng Ding, Zhijiang Guo, Michael Schlichtkrull, and Andreas Vlachos. AVerIMaTeC: A dataset for automatic verification of image-text claims with evidence from the web. In *Advances in Neural Information Processing Systems, Datasets and Benchmarks Track*, volume 38, pp. 4890–4923. Curran Associates, Inc., 2025. doi: 10.52202/085713-0154. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/0695cbd1f85f54780e2b351db4252a97-Abstract-Datasets_and_Benchmarks_Track.html.

Lin Chen, Jinsong Li, Xiaoyi Dong, et al. Are we on the right way for evaluating large vision-language models? In *Advances in Neural Information Processing Systems*, volume 37, pp. 27056–27087. Curran Associates, Inc., 2024. doi: 10.52202/079017-0850. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/2f8ee6a3d766b426d2618e555b5aeb39-Abstract-Conference.html.

Anthony G. Cohn and Jochen Renz. Qualitative spatial representation and reasoning. In Frank van Harmelen, Vladimir Lifschitz, and Bruce Porter (eds.), *Handbook of Knowledge Representation*, volume 3 of *Foundations of Artificial Intelligence*, pp. 551–596. Elsevier, 2008. doi: 10.1016/S1574-6526(07)03013-1. URL https://doi.org/10.1016/S1574-6526(07)03013-1.

Muhammad Sohail Danish, Muhammad Akhtar Munir, Syed Roshaan Ali Shah, Kartik Kuckreja, Fahad Shahbaz Khan, Paolo Fraccaro, Alexandre Lacoste, and Salman Khan. GEOBench-VLM: Benchmarking vision-language models for geospatial tasks. In *Proceedings of*

the IEEE/CVF International Conference on Computer Vision (ICCV), pp. 7132–7142, 2025. doi: 10.1109/ICCV51701.2025.00670. URL https://openaccess.thecvf.com/content/ICCV2025/html/Danish_GEOBench-VLM_Benchmarking_Vision-Language_Models_for_Geospatial_Tasks_ICCV_2025_paper.html.

Jay DeYoung, Sarthak Jain, Nazneen Fatema Rajani, Eric Lehman, Caiming Xiong, Richard Socher, and Byron C. Wallace. ERASER: A benchmark to evaluate rationalized NLP models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 4443–4458, Online, 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.408. URL https://aclanthology.org/2020.acl-main.408/.

Jian Ding, Nan Xue, Gui-Song Xia, et al. Object detection in aerial images: A large-scale benchmark and challenges. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44 (11):7778–7796, 2022. doi: 10.1109/TPAMI.2021.3117983. URL https://doi.org/10.1109/TPAMI.2021.3117983.

Bowen Dong, Minheng Ni, Zitong Huang, Guanglei Yang, Wangmeng Zuo, and Lei Zhang. MIRAGE: Assessing hallucination in multimodal reasoning chains of MLLM. In *Advances in Neural Information Processing Systems*, volume 38, pp. 136305–136350. Curran Associates, Inc., 2025. doi: 10.52202/085713-4099. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/b238324b309da12c7446d92c14db9f7e-Abstract-Conference.html.

Ronghao Fu, Haoran Liu, Weijie Zhang, Zhiwen Lin, Xiao Yang, Peng Zhang, and Bo Yang. OmniEarth: A benchmark for evaluating vision-language models in geospatial tasks. In *Proceedings of the 32nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.2*, pp. 8870–8881. Association for Computing Machinery, 2026. doi: 10.1145/3770855.3817467. URL https://doi.org/10.1145/3770855.3817467.

GLM-V Team, Wenyi Hong, Wenmeng Yu, et al. GLM-4.5V and GLM-4.1V-Thinking: Towards versatile multimodal reasoning with scalable reinforcement learning. Technical report, arXiv, 2025. URL https://arxiv.org/abs/2507.01006.

Google DeepMind. Gemini 3.1 Pro model card. Model card, 2026. URL https://deepmind.google/models/model-cards/gemini-3-1-pro.

Ritwik Gupta, Richard Hosfelt, Sandra Sajeev, Nirav Patel, Bryce Goodman, Jigar Doshi, Eric Heim, Howie Choset, and Matthew Gaston. xBD: A dataset for assessing building damage from satellite imagery. Technical report, arXiv, 2019. URL https://arxiv.org/abs/1911.09296.

Daniel Hernández. *Qualitative Representation of Spatial Knowledge*, volume 804 of *Lecture Notes in Computer Science*. Springer, 1994. doi: 10.1007/BFb0020328. URL https://doi.org/10.1007/BFb0020328.

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. 6693–6702, 2019. doi: 10.1109/CVPR.2019.00686. URL https://openaccess.thecvf.com/content_CVPR_2019/html/Hudson_GQA_A_New_Dataset_for_Real-World_Visual_Reasoning_and_Compositional_CVPR_2019_paper.html.

Alon Jacovi and Yoav Goldberg. Towards faithfully interpretable NLP systems: How should we define and evaluate faithfulness? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 4198–4205, Online, 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.386. URL https://aclanthology.org/2020.acl-main.386/.

Dongzhi Jiang, Renrui Zhang, Ziyu Guo, et al. MME-CoT: Benchmarking chain-of-thought in large multimodal models for reasoning quality, robustness, and efficiency. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 27793–27830. PMLR, 2025. URL https://proceedings.mlr.press/v267/jiang25n.html.

Liqiang Jing, Ruosen Li, Yunmo Chen, and Xinya Du. FaithScore: Fine-grained evaluations of hallucinations in large vision-language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 5042–5063, Miami, Florida, USA, 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-emnlp.290. URL https://aclanthology.org/2024.findings-emnlp.290/.

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. 1988–1997, 2017. doi: 10.1109/CVPR.2017.215. URL https://openaccess.thecvf.com/content_cvpr_2017/html/Johnson_CLEVR_A_Diagnostic_CVPR_2017_paper.html.

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. doi: 10.1109/CVPR52733.2024.02629. URL https://openaccess.thecvf.com/content/CVPR2024/html/Kuckreja_GeoChat_Grounded_Large_Vision-Language_Model_for_Remote_Sensing_CVPR_2024_paper.html.

Tamera Lanham, Anna Chen, Ansh Radhakrishnan, et al. Measuring faithfulness in chain-of-thought reasoning. Technical report, arXiv, 2023. URL https://arxiv.org/abs/2307.13702.

Bo Li, Yuanhan Zhang, Dong Guo, et al. LLaVA-OneVision: Easy visual task transfer. *Transactions on Machine Learning Research*, 2025. ISSN 2835-8856. URL https://openreview.net/forum?id=zKv8qULV6n.

Bohao Li, Yuying Ge, Yixiao Ge, Guangzhi Wang, Rui Wang, Ruimao Zhang, and Ying Shan. SEED-Bench-2: Benchmarking multimodal large language models. Technical report, arXiv, 2023a. URL https://arxiv.org/abs/2311.17092.

Ke Li, Gang Wan, Gong Cheng, Liqiu Meng, and Junwei Han. Object detection in optical remote sensing images: A survey and a new benchmark. *ISPRS Journal of Photogrammetry and Remote Sensing*, 159:296–307, 2020. doi: 10.1016/j.isprsjprs.2019.11.023. URL https://doi.org/10.1016/j.isprsjprs.2019.11.023.

Kun Li, George Vosselman, and Michael Ying Yang. HRVQA: A visual question answering benchmark for high-resolution aerial images. *ISPRS Journal of Photogrammetry and Remote Sensing*, 214:65–81, 2024a. doi: 10.1016/j.isprsjprs.2024.06.002. URL https://doi.org/10.1016/j.isprsjprs.2024.06.002.

Wenshuai Li, Xiantai Xiang, Zixiao Wen, Guangyao Zhou, Ben Niu, Feng Wang, Lijia Huang, Qiantong Wang, and Yuxin Hu. GeoReason: Aligning thinking and answering in remote sensing vision-language models via logical consistency reinforcement learning. Technical report, arXiv, 2026a. URL https://arxiv.org/abs/2601.04118.

Xiang Li, Jian Ding, and Mohamed Elhoseiny. VRSBench: A versatile vision-language benchmark dataset for remote sensing image understanding. In *Advances in Neural Information Processing Systems, Datasets and Benchmarks Track*, volume 37, pp. 3229–3242. Curran Associates, Inc., 2024b. doi: 10.52202/079017-0106. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/05b7f821234f66b78f99e7803fffa78a-Abstract-Datasets_and_Benchmarks_Track.html.

Xuchen Li, Xuzhao Li, Renjie Pi, Shiyu Hu, Jian Zhao, and Jiahui Gao. Beyond accuracy: Evaluating grounded visual evidence in thinking with images. Technical report, arXiv, 2026b. URL https://arxiv.org/abs/2601.11633.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. In *Proceedings of the 2023 Conference on*

Empirical Methods in Natural Language Processing, pp. 292–305, Singapore, 2023b. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.20. URL https://aclanthology.org/2023.emnlp-main.20/.

Jiaqi Liu, Lang Sun, Ronghao Fu, and Bo Yang. Towards faithful reasoning in remote sensing: A perceptually-grounded geospatial chain-of-thought for vision-language models. In *International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=lJ7zecny2e.

Yuan Liu, Haodong Duan, Yuanhan Zhang, et al. MMBench: Is your multi-modal model an all-around player? In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 216–233. Springer Nature Switzerland, 2024. doi: 10.1007/978-3-031-72658-3_13. URL https://doi.org/10.1007/978-3-031-72658-3_13.

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. URL https://doi.org/10.1109/TGRS.2020.2988782.

Zhiming Luo, Di Wang, Haonan Guo, Jing Zhang, and Bo Du. VLRS-Bench: A vision-language reasoning benchmark for remote sensing. Technical report, arXiv, 2026. URL https://arxiv.org/abs/2602.07045.

Weijiang Lv, Yaoxuan Feng, Xiaobo Xia, Jiayu Wang, Yan Jing, Wenchao Chen, and Bo Chen. SPD-Faith Bench: Diagnosing and improving faithfulness in chain-of-thought for multimodal large language models. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 19875–19927, San Diego, California, United States, 2026. Association for Computational Linguistics. doi: 10.18653/v1/2026.findings-acl.995. URL https://aclanthology.org/2026.findings-acl.995/.

Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. On faithfulness and factuality in abstractive summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 1906–1919, Online, 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.173. URL https://aclanthology.org/2020.acl-main.173/.

OpenAI. GPT-4o system card. Technical report, arXiv, 2024. URL https://arxiv.org/abs/2410.21276.

OpenAI. GPT-5.1 Instant and GPT-5.1 Thinking system card addendum. System card, 2025. URL https://openai.com/index/gpt-5-system-card-addendum-gpt-5-1/.

Chao Pang, Xingxing Weng, Jiang Wu, et al. VHM: Versatile and honest vision language model for remote sensing image analysis. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(6):6381–6388, 2025. doi: 10.1609/aaai.v39i6.32683. URL https://ojs.aaai.org/index.php/AAAI/article/view/32683.

Qwen Team. Qwen3.5: Towards native multimodal agents. Blog post, 2026. URL https://qwen.ai/blog?id=qwen3.5.

Shreelekha Revankar, Utkarsh Mall, Cheng Perng Phoo, Kavita Bala, and Bharath Hariharan. MONITRS: Multimodal observations of natural incidents through remote sensing. In *Advances in Neural Information Processing Systems, Datasets and Benchmarks Track*, volume 38, pp. 144930–144955. Curran Associates, Inc., 2025. doi: 10.52202/085713-4355. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/bd96a50dfd2314e48787581840a07a1a-Abstract-Datasets_and_Benchmarks_Track.html.

Anna Rohrbach, Lisa Anne Hendricks, Kaylee Burns, Trevor Darrell, and Kate Saenko. Object hallucination in image captioning. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 4035–4045, Brussels, Belgium, 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1437. URL https://aclanthology.org/D18-1437/.

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. URL https://proceedings.mlr.press/v267/shabbir25a.html.

Run Shao, Ziyu Li, Zhaoyang Zhang, et al. Asking like Socrates: Socrates helps VLMs understand remote sensing images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 26465–26475, 2026. URL https://openaccess.thecvf.com/content/CVPR2026/html/Shao_Asking_like_Socrates_Socrates_helps_VLMs_understand_remote_sensing_images_CVPR_2026_paper.html.

Li Shen, Yao Lu, Hao Chen, Hao Wei, Donghai Xie, Jiabao Yue, Rui Chen, Shouye Lv, and Bitao Jiang. S2Looking: A satellite side-looking dataset for building change detection. *Remote Sensing*, 13(24):5094, 2021. doi: 10.3390/rs13245094. URL https://doi.org/10.3390/rs13245094.

Yan Shu, Bin Ren, Zhitong Xiong, Xiao Xiang Zhu, Begüm Demir, Nicu Sebe, and Paolo Rota. TerraScope: Pixel-grounded visual reasoning for Earth observation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 16712–16722, 2026. URL https://openaccess.thecvf.com/content/CVPR2026/html/Shu_TerraScope_Pixel-Grounded_Visual_Reasoning_for_Earth_Observation_CVPR_2026_paper.html.

Ashbindu Singh. Digital change detection techniques using remotely-sensed data. *International Journal of Remote Sensing*, 10(6):989–1003, 1989. doi: 10.1080/01431168908903939. URL https://doi.org/10.1080/01431168908903939.

Xian Sun, Peijin Wang, Zhiyuan Yan, et al. FAIR1M: A benchmark dataset for fine-grained object recognition in high-resolution remote sensing imagery. *ISPRS Journal of Photogrammetry and Remote Sensing*, 184:116–130, 2022. doi: 10.1016/j.isprsjprs.2021.12.004. URL https://doi.org/10.1016/j.isprsjprs.2021.12.004.

Sree Harsha Tanneru, Dan Ley, Chirag Agarwal, and Himabindu Lakkaraju. On the hardness of faithful chain-of-thought reasoning in large language models. Technical report, arXiv, 2024. URL https://arxiv.org/abs/2406.10625.

Mai Tsujimoto, Junjue Wang, Weihao Xuan, and Naoto Yokoya. Geo3DVQA: Evaluating vision-language models for 3D geospatial reasoning from aerial imagery. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 4994–5003, 2026. doi: 10.1109/WACV61042.2026.00485. URL https://openaccess.thecvf.com/content/WACV2026/html/Tsujimoto_Geo3DVQA_Evaluating_Vision-Language_Models_for_3D_Geospatial_Reasoning_from_Aerial_WACV_2026_paper.html.

Miles Turpin, Julian Michael, Ethan Perez, and Samuel R. Bowman. Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting. In *Advances in Neural Information Processing Systems*, volume 36, pp. 74952–74965. Curran Associates, Inc., 2023. doi: 10.52202/075280-3275. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/ed3fea9033a80fea1376299fa7863f4a-Abstract-Conference.html.

Di Wang, Shunyu Liu, Wentao Jiang, et al. GeoZero: Incentivizing reasoning from scratch on geospatial scenes. Technical report, arXiv, 2025a. URL https://arxiv.org/abs/2511.22645.

Fengxiang Wang, Hongzhen Wang, Zonghao Guo, et al. XLRS-Bench: Could your multimodal LLMs understand extremely large ultra-high-resolution remote sensing imagery? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14325–14336, 2025b. doi: 10.1109/CVPR52734.2025.01336. URL https://openaccess.thecvf.com/content/CVPR2025/html/Wang_XLRS-Bench_Could_Your_Multimodal_LLMs_Understand_Extremely_Large_Ultra-High-Resolution_Remote_CVPR_2025_paper.html.

Fengxiang Wang, Mingshuo Chen, Yueying Li, et al. GeoEyes: On-demand visual focusing for evidence-grounded understanding of ultra-high-resolution remote sensing imagery. Technical report, arXiv, 2026a. URL https://arxiv.org/abs/2602.14201.

Haochen Wang, Xiangtai Li, Zilong Huang, et al. Traceable evidence enhanced visual grounded reasoning: Evaluation and method. In *International Conference on Learning Representations*, 2026b. URL https://openreview.net/forum?id=zwMpimSZEq.

Junjue Wang, Zhuo Zheng, Zihang Chen, Ailong Ma, and Yanfei Zhong. EarthVQA: Towards queryable Earth via relational reasoning-based remote sensing visual question answering. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(6):5481–5489, 2024. doi: 10.1609/aaai.v38i6.28357. URL https://ojs.aaai.org/index.php/AAAI/article/view/28357.

Junjue Wang, Weihao Xuan, Heli Qi, et al. DisasterM3: A remote sensing vision-language dataset for disaster damage assessment and response. In *Advances in Neural Information Processing Systems, Datasets and Benchmarks Track*, volume 38, pp. 179475–179489. Curran Associates, Inc., 2025c. doi: 10.52202/085713-5398. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/ec80d18205e39d42a27192d5f3ddd688-Abstract-Datasets_and_Benchmarks_Track.html.

Junyang Wang, Yuhang Wang, Guohai Xu, et al. AMBER: An LLM-free multi-dimensional benchmark for MLLMs hallucination evaluation. Technical report, arXiv, 2023. URL https://arxiv.org/abs/2311.07397.

Weiyun Wang, Zhangwei Gao, Lixin Gu, et al. InternVL3.5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. Technical report, arXiv, 2025d. URL https://arxiv.org/abs/2508.18265.

Zelin Xu, Yupu Zhang, Saugat Adhikari, Saiful Islam, Tingsong Xiao, Zibo Liu, Shigang Chen, Da Yan, and Zhe Jiang. EarthSpatialBench: Benchmarking spatial reasoning capabilities of multimodal LLMs on Earth imagery. Technical report, arXiv, 2026. URL https://arxiv.org/abs/2602.15918.

Weihao Xuan, Junjue Wang, Heli Qi, Zihang Chen, Zhuo Zheng, Yanfei Zhong, Junshi Xia, and Naoto Yokoya. DynamicVL: Benchmarking multimodal large language models for dynamic city understanding. In *Advances in Neural Information Processing Systems, Datasets and Benchmarks Track*, volume 38, pp. 37321–37339. Curran Associates, Inc., 2025. doi: 10.52202/085713-1114. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/2fb80f455dc6056422c8af87dca50489-Abstract-Datasets_and_Benchmarks_Track.html.

Qiwei Yan, Chubo Deng, Chenglong Liu, et al. ReCon1M: A large-scale benchmark dataset for relation comprehension in remote sensing imagery. *IEEE Transactions on Geoscience and Remote Sensing*, 63:1–22, 2025. doi: 10.1109/TGRS.2025.3589986. URL https://doi.org/10.1109/TGRS.2025.3589986.

An Yang, Anfeng Li, Baosong Yang, et al. Qwen3 technical report. Technical report, arXiv, 2025. URL https://arxiv.org/abs/2505.09388.

Kunping Yang, Gui-Song Xia, Zicheng Liu, Bo Du, Wen Yang, Marcello Pelillo, and Liangpei Zhang. Asymmetric Siamese networks for semantic change detection in aerial images. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–18, 2022. doi: 10.1109/TGRS.2021.3113912. URL https://doi.org/10.1109/TGRS.2021.3113912.

Liang Yao, Fan Liu, Hongbo Lu, Chuanyi Zhang, Rui Min, Shengxiang Xu, Shimin Di, and Pai Peng. RemoteReasoner: Towards unifying geospatial reasoning workflow. *Proceedings of the AAAI Conference on Artificial Intelligence*, 40(14):11883–11891, 2026. doi: 10.1609/aaai.v40i14.38175. URL https://ojs.aaai.org/index.php/AAAI/article/view/38175.

Zhihan Yin, Jianxin Liang, Yueqian Wang, Yifeng Yao, Huishuai Zhang, and Dongyan Zhao. FREAK: A fine-grained hallucination evaluation benchmark for advanced MLLMs. In *International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=YeagC09j2K.

Wei Zhang, Miaoxin Cai, Tong Zhang, Yin Zhuang, and Xuerui Mao. EarthGPT: A universal multimodal large language model for multisensor image comprehension in remote sensing domain. *IEEE Transactions on Geoscience and Remote Sensing*, 62:1–20, 2024. doi: 10.1109/TGRS.2024.3409624. URL https://doi.org/10.1109/TGRS.2024.3409624.

Zihui Zhou, Yong Feng, Yanying Chen, Guofan Duan, Zhenxi Song, Mingliang Zhou, and Weijia Jia. RSHallu: Dual-mode hallucination evaluation for remote-sensing multimodal large language models with domain-tailored mitigation. Technical report, arXiv, 2026. URL https://arxiv.org/abs/2602.10799.

Xing Zi, Jinghao Xiao, Yunxiao Shi, Xian Tao, Jun Li, Ali Braytee, and Mukesh Prasad. RSVLM-QA: A benchmark dataset for remote sensing vision language model-based question answering. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pp. 12905–12911. Association for Computing Machinery, 2025. doi: 10.1145/3746027.3758235. URL https://doi.org/10.1145/3746027.3758235.

Appendix

A Construction Details **18**
A.1 Data Sources . . . 18
A.2 Question Construction . . . 19
A.3 Evidence Scope . . . 20
A.4 Question Validation . . . 20
B Evaluation Details **21**
B.1 Model Setup . . . 21
B.2 Prompting and Parsing . . . 22
B.3 Claim Verification . . . 24
B.4 Metrics . . . 25
C Detailed Results **25**
C.1 Subtype Results . . . 26
C.2 Model Claim Profiles . . . 26
C.3 Design Properties of FA . . . 28
D Quality Control **28**
D.1 Human Audit . . . 29
D.2 Contradiction Audit . . . 29
D.3 Human Trust Study . . . 30
D.4 Visual Dependence . . . 31
D.5 Extractor Quality . . . 31
D.6 Extractor Robustness . . . 31
D.7 Verification Coverage . . . 31
D.8 Prompt Sensitivity . . . 32
D.9 Fixed Evidence Budget . . . 33
D.10 Answer Format . . . 33
E Case Studies **34**
E.1 Representative Cases . . . 34
E.2 Failure Types . . . 34
F Limitations and Broader Impact **35**

A Construction Details

This section describes the source annotations, question programs, evidence scopes, and validation process used to construct RSFaith-Bench.

Table A1: Comparison with representative general, faithfulness, and remote sensing benchmarks across data sources, question counts, annotation method, evaluation target, and two properties of the evaluation. Suff.: the stated visual facts are checked for sufficiency with respect to the answer. Refut.: a stated claim can be refuted against source annotation (△: object presence or edited image differences only). ○, ✓, and ⊘ denote automatic generation, human authoring, and machine-assisted annotation with human verification, respectively.

Benchmark	Year	Data Sources	# Questions	Annotation Method	Evaluation Target	Suff.	Refut.
General Benchmarks							
MMBench (Liu et al., 2024)	2024	10 Public Datasets	3,217	⊘	Answer	✗	✗
MMStar (Chen et al., 2024)	2024	6 Public Benchmarks	1,500	✓	Answer	✗	✗
SEED-Bench-2 (Li et al., 2023a)	2023	5 Public Datasets	24,000	⊘	Answer	✗	✗
Faithfulness and Hallucination Benchmarks							
POPE (Li et al., 2023b)	2023	COCO	9,000	○	Object presence	✗	△
AMBER (Wang et al., 2023)	2023	Natural images	15,220	✓	Claim truth	✗	△
FaithScore (Jing et al., 2024)	2024	LLaVA-1k, COCO	Metric	○	Atomic claim truth	✗	✗
MME-CoT (Jiang et al., 2025)	2025	6 domains	1,130	✓	Reasoning-trace quality	✗	✗
TreeBench (Wang et al., 2026b)	2025	Natural images	405	✓	Answer / Evidence region	✗	✗
ViEBench (Li et al., 2026b)	2026	Web, VisualProbe	200	✓	Answer / Evidence region	✗	✗
SPD-Faith (Lv et al., 2026)	2026	COCO (edited pairs)	3,000	⊘	CoT–evidence alignment	✗	△
RSHallu (Zhou et al., 2026)	2026	4 RS datasets	2,023	✓	Hallucination (model-judged)	✗	✗
Remote Sensing Benchmarks							
RSVQA (Lobry et al., 2020)	2020	HR & LR Datasets	1M+	○	Answer	✗	✗
HRVQA (Li et al., 2024a)	2024	Aerial imagery	1,070,240	○	Answer	✗	✗
EarthVQA (Wang et al., 2024)	2024	LoveDA Dataset	208,593	⊘	Answer / Mask	✗	✗
VRSBench (Li et al., 2024b)	2024	DIOR, DOTA-v2	123,221	⊘	Answer / Region	✗	✗
DisasterM3 (Wang et al., 2025c)	2025	Bi-temporal disasters	123K	⊘	Answer / Mask	✗	✗
XLRS-Bench (Wang et al., 2025b)	2025	6 UHR RS Datasets	32K	✓	Answer / Region	✗	✗
RSVLM-QA (Zi et al., 2025)	2025	4 public RS datasets	162,373	○	Answer	✗	✗
VLRS-Bench (Luo et al., 2026)	2026	11 Public Datasets	2,000	⊘	Answer	✗	✗
OmniEarth (Fu et al., 2026)	2026	High-res EO imagery	44,210	⊘	Answer / Region / Mask	✗	✗
RSFaith-Bench	**2026**	**6 public RS datasets**	**13,511**	⊘	**Answer + Stated Evidence**	✓	✓

A.1 Data Sources

RSFaith-Bench is built from six public remote sensing datasets with complementary annotation regimes. Every question is generated from the held-out evaluation split of its source dataset, disjoint from the split each dataset provides for training. DIOR, DOTA-v2.0, and FAIR1M provide object-centric supervision for single-image perception questions; SECOND and xBD provide bi-temporal annotations for temporal reasoning; and ReCon1M provides explicit relation triplets for relational reasoning.

DIOR. DIOR (Li et al., 2020) contains 23,463 optical remote sensing images and 192,472 annotated instances from 20 object categories. The dataset covers diverse object scales, scene types, seasons, weather conditions, and imaging quality. In RSFaith-Bench, DIOR is mainly used for single-image object facts such as existence, counting, and coarse location.

DOTA-v2.0. DOTA-v2.0 (Ding et al., 2022) contains 11,268 aerial images, 1,793,658 annotated instances, and 18 object categories. Its oriented bounding boxes support geometry-sensitive verification when objects are densely packed or appear at arbitrary rotations. RSFaith-Bench uses DOTA-v2.0 for object presence, localization, counting, and spatial-anchor facts.

FAIR1M. FAIR1M (Sun et al., 2022) targets fine-grained object recognition in high-resolution remote sensing imagery. It combines images from Gaofen satellites and Google Earth at sub-meter

Table A2: Fact types used in the materialized fact layer $F(G)$.

Fact type	Canonical form	Primary provenance	Completeness requirement
Existence	$\textsc{exist}(c, a, t)$	Object boxes, masks, regions	Category c annotated within spatial anchor a and time t
Counting	$\textsc{count}(c, a, t) = n$	Instance annotations or semantic components	Exhaustive annotation for category c in the queried scope
Attribute	$\textsc{attr}(u, h, t) = z$	Class labels, damage states, region states	Attribute channel h available for the referred entity or region at time t
Location	$\textsc{loc}(u, a)$	Native geometry and derived anchors	Geometry readable and anchor definition fixed
Relation	$\textsc{rel}(u, r, v, \tau)$	Relation triplets, deterministic geometry, correspondences	Predicate r closed or derivable in the queried scope

resolution and annotates objects with oriented bounding boxes across coarse categories and fine-grained subcategories. RSFaith-Bench uses FAIR1M for fine-grained recognition, subtype evidence, category-specific counting, and location evidence.

SECOND. SECOND (Yang et al., 2022) provides 4,662 pairs of multi-temporal aerial images from urban areas including Hangzhou, Chengdu, and Shanghai. Each pair is divided into 512×512 patches with pixel-level semantic labels over six land-cover classes. The dataset supports time-anchored region states, semantic transitions, and net-change evidence over land-cover classes.

xBD. xBD (Gupta et al., 2019) provides pre- and post-event satellite imagery with building annotations over disaster-affected areas. The original release contains 850,736 building annotations collected over 45,362 square kilometers, with each building instance associated with a damage-level label. RSFaith-Bench uses xBD for temporal questions involving building state changes and post-event conditions.

ReCon1M. ReCon1M (Yan et al., 2025) focuses on relation comprehension in remote sensing imagery. The released benchmark includes 21,392 images, 859,751 object boxes spanning 60 categories, and 1,149,342 relation triplets over 64 relation types. It is the primary source for directional, topological, proximity, projective-ordering, and aggregate-distribution questions.

A.2 Question Construction

Each sample is represented by a grounded scene graph $G = (V, E, A)$. Nodes in V correspond to object instances, semantic regions, building footprints, or change units. Edges in E encode spatial relations within a time point and correspondences across time. Attributes in A store canonical labels, geometry, temporal anchors, damage states, and other typed fields derived from source annotations. Every node, edge, and attribute keeps provenance links to the original annotation unit.

The graph is materialized as a finite fact layer $F(G)$. Question generation, claim verification, and answer reconstruction all operate on this shared interface. Table A2 summarizes the fact types used by the benchmark.

Each question is generated by a program $P_q = (R_q, f_q)$. The canonical evidence schema R_q defines the typed facts consumed by the question program, and the answer function f_q maps a completed slot assignment to the gold answer. The benchmark covers unary perception programs, entity-relation programs, and temporal programs. Perception programs use existence, count, attribute, or location slots; relational programs additionally bind compatible subject-object pairs and relation predicates; temporal programs bind states or comparable counts at two time anchors, with correspondence evidence when the answer requires cross-time matching. Figure 8 shows one rendered question for each of the 12 subtypes.

Question instances are filtered before natural-language rendering. A program is instantiated only when all slots in R_q are filled by facts from closed scopes (Appendix A.3). Ambiguous references, borderline geometry predicates, unresolved temporal correspondences, and templates with multiple equally valid answers are removed before evaluation.

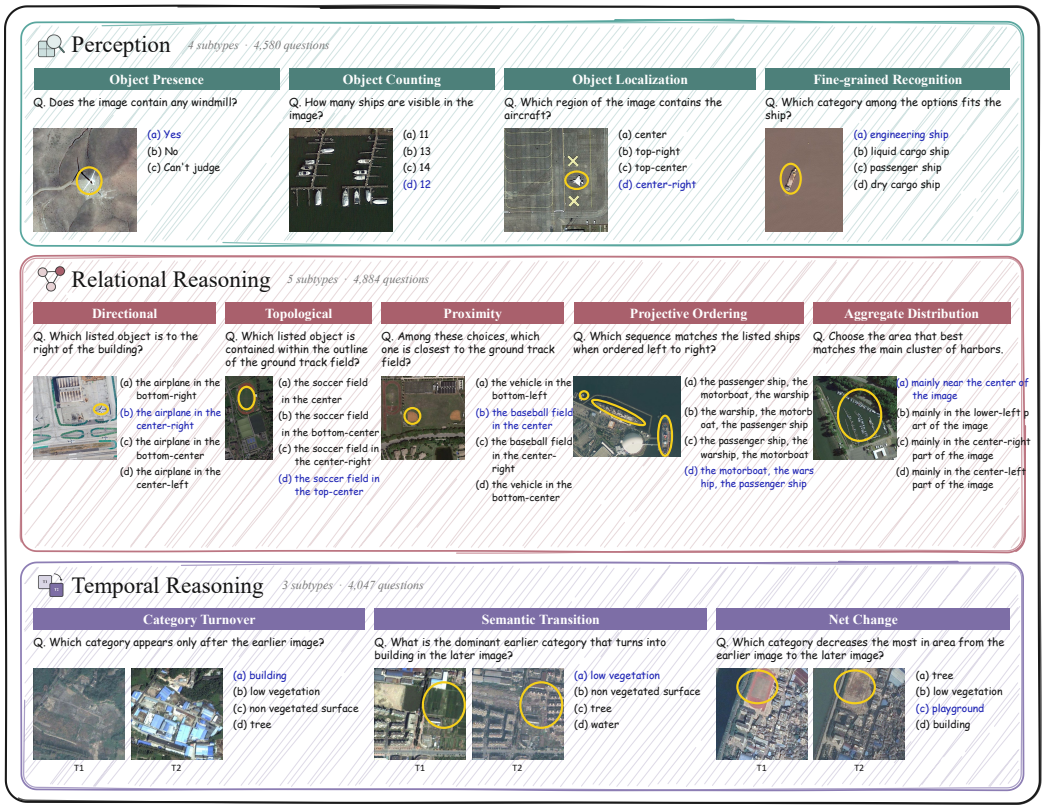


Figure 8: Example RSFaith-Bench questions, one per subtype, grouped by reasoning level. Each example shows the image (or the bi-temporal pair T1/T2), the question, and the answer options; the gold option is shown in blue, and yellow ellipses mark the objects or regions that determine the answer.

Worked Example. For a counting question such as "How many aircraft are visible in the airport region?", the program binds a closed counting scope for aircraft inside the queried region and creates a required COUNTING slot. The answer function maps the count value from $F(G)$ to the rendered answer option. At evaluation time, a model claim asserting the same exact count can fill this slot; a different exact count is verified as contradicted within the closed scope.

A.3 EVIDENCE SCOPE

A *closed scope* $\Omega = (p, \sigma, a, t)$ is recorded on evidence slots whose corresponding fact inventory is complete. A claim is supported when it matches a fact in $F(G)$; contradiction, absence, and exact-count disagreement are assigned inside recorded closed scopes, and all other claims remain `uncertain`. Object datasets provide complete scopes for annotated categories and geometry-derived properties, temporal datasets for labeled time-anchored states and transitions, and ReCon1M for its declared relation predicates.

A.4 QUESTION VALIDATION

Protocol. We engage three annotators with graduate-level training in remote sensing or computer vision. Annotators first calibrate on a stratified pilot of 200 candidate questions spanning the twelve question subtypes; disagreements are discussed jointly to align judgments under the validation protocol below. Inter-annotator agreement is then measured on a separate, multiply annotated held-out set of 100 questions, yielding $\kappa = 0.81$. Pilot questions are not included in the released benchmark. The remaining candidate questions are stratified by source dataset, reasoning level, question subtype, and evidence family, and each candidate is randomly assigned to one annotator within its stratum for full validation. The assigned annotator jointly checks the question, gold answer y_q, and evidence slots R_q against the source image or image pair and its annotation. Local wording, cat-

egory normalization, or option-formatting errors are corrected when the intended visual evidence remains unchanged; questions with unresolved ambiguity, self-contained answers, or insufficient visual support are removed. Validation of the 13,511 released questions requires approximately 337 annotator-hours.

Guidelines. Given a candidate instance consisting of the source image or image pair, the rendered question, the gold answer y_q, the reference evidence slots R_q, and the underlying fact record $F(G)$, the reviewer verifies whether the question has a unique visually grounded answer, whether the gold answer follows from $F(G)$, and whether every required slot in R_q is instantiable from facts in $F(G)$ within a closed scope. The question, answer, image evidence, and fact record are checked as a single unit rather than as independent fields. Modification is preferred over removal when the issue is local and the intended visual evidence remains unchanged; questions are removed when the source evidence is incomplete or the answer depends on subjective interpretation.

General validity. A question is accepted only when its entity references, spatial anchors, temporal anchors, and answer options are unambiguous. Questions are removed if (i) more than one answer option is consistent with both the image and the source annotation, (ii) the queried object or region is not fully contained within the image, or (iii) the gold answer would change if the reviewer relied solely on the image versus solely on the source annotation.

Perception questions. For object existence, counting, localization, and fine-grained recognition questions, the annotator verifies that the asserted category, count, attribute, or location agrees with both the image content and the source annotation. Counting questions are removed when the reviewer cannot reproduce the gold count from the image without consulting the source annotation, or when the source annotation is itself ambiguous about which instances fall within the queried region.

Relational questions. For relational questions, the subject and object must each be uniquely identifiable from the question text and the image; pairs are removed when more than one image entity satisfies the subject or object description. The relation predicate is checked against both the image and the relation annotation, and the question is removed if the reviewer's verdict differs between the two.

Temporal questions. For temporal questions, both time anchors are verified separately against the corresponding images and annotations. Change, persistence, or net-difference questions are accepted only when the pre-state and post-state can each be determined independently from the corresponding image and annotation, and when any required cross-time correspondence is recorded in the source annotation rather than inferred by the annotator.

B Evaluation Details

B.1 Model Setup

Table A3 provides details of all open-source MLLMs evaluated in RSFaith-Bench. All models use the shared evidence-request prompt in Appendix B.2. Open-source MLLMs are run locally on 8 NVIDIA H100 GPUs (80 GB each) with each model's default generation configuration. Proprietary MLLMs are accessed through their official APIs, using the following snapshots: GPT-4o (`gpt-4o-2024-11-20`), GPT-5.1 (`gpt-5.1-2025-11-13`), Claude Haiku 4.5 (`claude-haiku-4-5-20251001`), Claude Sonnet 4.6 (`claude-sonnet-4-6-20260217`), and Gemini 3.1 Pro (`gemini-3.1-pro-preview`). Each model is queried through its standard response interface with the evidence-request prompt, and the text it returns is scored as its response. Decoding parameters are fixed where the interface exposes them and otherwise left at the interface default. Automated scores on the stratified 500-question subset are averaged over two scoring runs.

Remote sensing reasoning models. GeoReason (Li et al., 2026a), GeoZero-8B (Wang et al., 2025a), RemoteReasoner (Yao et al., 2026), and RS-EoT-7B (Shao et al., 2026) are run from their public checkpoints with their official inference settings.

Table A3: Model configurations for the evaluated open-source MLLMs. For remote sensing models fine-tuned from a general MLLM, the language-backbone column gives the base MLLM and its training stages.

Model	Vision Backbone	Language Backbone	Params
General-Domain MLLMs			
Qwen2.5-VL-7B	Dynamic-resolution ViT	Qwen2.5-7B	7B
Qwen2.5-VL-72B	Dynamic-resolution ViT	Qwen2.5-72B	72B
Qwen3-VL-8B	SigLIP-2	Qwen3-8B	8B
Qwen3-VL-32B	SigLIP-2	Qwen3-32B	32B
Qwen3.5-9B	Built-in vision encoder	Gated DeltaNet hybrid	9B
Qwen3.5-27B	Built-in vision encoder	Gated DeltaNet hybrid	27B
InternVL3.5-8B	InternViT-300M	Qwen3-8B	8B
InternVL3.5-38B	InternViT-6B	Qwen3-32B	38B
LLaVA-OneVision-7B	SigLIP SO400M	Qwen2-7B	7B
GLM-4.5V	AIMv2-Huge ViT	GLM-4.5-Air (MoE)	106B (12B-A)
Remote Sensing MLLMs			
GeoChat	CLIP ViT-L/14-336px	Vicuna-v1.5-7B	7B
VHM	CLIP ViT-L/14-336px	Vicuna-v1.5-7B	7B
RSThinker	AIMv2-Huge ViT	GLM-4.1V-9B (SFT+GRPO)	10B
GeoReason	Dynamic-resolution ViT	Qwen2.5-VL-7B (SFT+GRPO)	7B
GeoZero-8B	SigLIP-2	Qwen3-VL-8B (SFT+A^{2}GRPO)	8B
RemoteReasoner	Dynamic-resolution ViT	Qwen2.5-VL-7B (GRPO)	7B
RS-EoT-7B	Dynamic-resolution ViT	Qwen2.5-VL-7B (SFT+GRPO)	7B

B.2 Prompting and Parsing

The prompt given to evaluated models is fixed across systems:

Model Response

Question: `{question}`
Image: `[image]`

Look at the image carefully. First describe the visual evidence you observe in the image that is relevant to this question. Based on this evidence, then state your answer. Begin with "Evidence:" and end with "Answer:".

The answer parser maps the stated answer $\hat{y}_q$ into the answer space of the current question program. Unmappable answers are scored as incorrect, while their evidence text is still processed for claim-level quantities. The evidence extractor is fixed across all evaluated systems. It first separates visual perception spans from analytical text, then decomposes visual spans into atomic typed claims and normalizes categories, attributes, relations, anchors, counts, polarity, and temporal references to the benchmark vocabulary. The label space supplied to the extractor contains only benchmark-supported category, relation, attribute, and location labels. Claims that cannot be stably mapped to this schema or label space are reported as unmapped normalization failures; they are not assigned verifier labels and do not enter CP.

Visual Span Identification

You are separating visual perception from analytical reasoning for schema-constrained claim extraction.

Task: Given a model rationale, process one focused sentence at a time. The input contains the previous sentence, the focused sentence, and the next sentence. Segment only the text inside `<FOCUS>...</FOCUS>` into short meaningful spans and label each span as `visual_perception` or `analytical`.

Label definitions:

- `visual_perception`: a statement about what is directly visible in the image, or directly comparable between earlier and later images. This includes object presence or absence, count, attribute, location, spatial relation, subtype/category classification grounded in visible appearance, and temporal visual change.
- `analytical`: interpretation, explanation, causal reasoning, answer selection, uncertainty, meta reasoning, or content that is not itself a direct visual statement.

Instructions:

- Only label and extract spans from text inside `<FOCUS>...</FOCUS>`. Use the surrounding context only to resolve pronouns, ellipsis, and time anchors.
- If a pronoun or phrase such as "it", "they", "one", "the other", or "this area" has a clear referent, rewrite it in `resolved_span`. Otherwise keep the original text and label it as `analytical`.
- Split mixed content whenever possible, but keep each span schema-complete. Preserve time anchors, negation, and scope-limiting clauses with the visual fact they modify.
- If visual cues identify an object's visible subtype or category, extract the subtype/category statement as `visual_perception` even when introduced by words such as "indicates", "suggests", "appears to be", "characteristic of", or "therefore".
- Do not include answer-selection phrases such as "therefore the answer is B" as visual evidence.

Output format: Return JSON only:

```
[{"span":"...", "resolved_span":"...",
  "label":"visual_perception|analytical"}]
```

Example 1:

```
Input:
  Context before: There is a ship next to the harbor.
  <FOCUS> It is located on the left side of the image, so option B is
    correct. </FOCUS>
  Context after:
Output:
[
  {"span":"It is located on the left side of the image",
   "resolved_span":
     "The ship is located on the left side of the image",
   "label":"visual_perception"},
  {"span":"so option B is correct",
   "resolved_span":"so option B is correct",
   "label":"analytical"}
]
```

Example 2:

```
Input:
  Context before: The deck is covered with pipes and manifolds.
  <FOCUS>This deck structure is characteristic of a liquid cargo ship,
     so the answer is A.</FOCUS>
  Context after:
Output:
[
  {"span":
     "This deck structure is characteristic of a liquid cargo ship",
   "resolved_span":"The ship is a liquid cargo ship",
   "label":"visual_perception"},
  {"span":"so the answer is A",
   "resolved_span":"so the answer is A",
   "label":"analytical"}
]
```

Atomic Claim Extraction

You are extracting atomic visual claims under a strict target schema.

Task: Convert each visual span into normalized atomic visual claims of types: Existence, Counting, Attribute, Relation, Location.

Rules:

- Each claim contains exactly one directly visually verifiable fact; split multiple facts into separate claims.
- Exclude reasoning, interpretation, uncertainty, and subjective language.
- Use only labels from the provided label space; subject and object are singular canonical nouns.
- Use `t1` for the single or earlier image and `t2` for the later image; decompose paired-image statements into separate time-anchored claims.
- Do not invent entities, times, relations, attributes, or locations; return `[]` if a statement cannot be normalized.
- Negative `Existence` applies only to image-global absence; encode regional absence as a negative `Location` claim.
- Do not convert vague quantities (e.g., "several", "many", "multiple") into exact counts.
- Encode relative spatial phrases (e.g., "to the right of X") as `Relation` claims, not Location.
- Encode subtype statements as `Attribute` claims with `attribute="subtype"`.

Schema: Return a JSON array only.

```
Existence:         {"type":"Existence","time":"t1|t2","subject":"...",
                    "value":true|false}
Counting (exact): {"type":"Counting","time":"t1|t2","subject":"...",
                    "quantity_kind":"exact","quantity":number}
Counting (cmp):    {"type":"Counting","time":"t1|t2","subject":"...",
                    "quantity_kind":"bound",
                    "quantity_operator":"gt|lt","object":"..."}
Attribute:         {"type":"Attribute","time":"t1|t2","subject":"...",
                    "attribute":"...","value":"..."}
Relation:          {"type":"Relation","time":"t1|t2","subject":"...",
                    "relation":"...","object":"...","value":true}
Pair transfer:     {"type":"Relation","time":"pair","subject":"...",
                    "relation":"transferred_to","object":"...",
                    "value":true}
Location:          {"type":"Location","time":"t1|t2","subject":"...",
                    "location":"...","value":true|false}
```

Example:

```
Input:  The ship in the center-left is a dry cargo ship.
Output: [
  {"type":"Existence","time":"t1","subject":"ship","value":true},
  {"type":"Location","time":"t1","subject":"ship",
   "location":"center-left","value":true},
  {"type":"Attribute","time":"t1","subject":"ship",
   "attribute":"subtype","value":"dry_cargo_ship"}
]
```

B.3 Claim Verification

Each mappable normalized claim $\bar{c}$ is verified against $F(G)$ with a three-valued evidence rule. Let $W^+(\bar{c}, F(G))$ indicate that the fact layer contains a positive witness for the claim, $W^-(\bar{c}, F(G))$ indicate that the observed value in a closed scope conflicts with the claim, and $K(\bar{c}, F(G))$ indicate that the relevant evidence channel is available. The verifier returns `support` if W^+ holds, `contradict` if W^+ is false while W^- and K hold, and `uncertain` otherwise.

Supported claims enter E_q^+, contradicted claims enter E_q^-, and uncertain claims enter E_q^u. Slot filling uses the normalized fields of E_q^+ to bind evidence slots in R_q. A claim fills a slot when its type, temporal anchor, semantic labels, predicate or attribute, expected value, and denotation are compatible with the slot definition, and a counting slot is also filled when supported counts over disjoint subregions covering its scope sum to its value. Contradicted, uncertain, and unmapped claims never fill slots.

Table A4: Verification rules for normalized visual claims.

Claim type	Positive witness W^+	Contradiction condition	Uncertain condition
Existence	Matched node set $V_s^{\tau,\alpha}$ agrees with polarity	Category is complete and observed existence disagrees	Category or scope is outside coverage
Counting	Observed count satisfies z and $\oplus$	Complete count scope gives a different count or comparison	Count value, operator, or category cannot be mapped
Attribute	A matched entity has attribute value z	Attribute channel exists and all observed values disagree	Attribute channel or referred entity is unresolved
Location	Geometry satisfies the anchor or point predicate	Geometry is readable and all candidates violate the location predicate	Geometry is missing or anchor is unstable
Relation	Explicit edge or stable geometry supports r	Predicate is closed and no candidate pair supports r	Predicate is open-world or candidate grounding is ambiguous

Algorithm 1 Verification of Visual Support

Require: sample (x, q, y) with fact layer $F(G)$, question program $P_q = (R_q, f_q)$, extractor vocabulary $\mathcal{O}$, model M

$(\hat{y}, t) \leftarrow M(x, q)$
$C \leftarrow \textsc{ExtractClaims}(t)$
$\bar{C} \leftarrow \emptyset$, $E_q^+ \leftarrow \emptyset$, $E_q^- \leftarrow \emptyset$, $E_q^u \leftarrow \emptyset$, $U_q \leftarrow \emptyset$
for each $c \in C$ **do**
  $\bar{c} \leftarrow \textsc{Normalize}(c, \mathcal{O})$
  if $\bar{c}$ is not stably mappable **then**
    $U_q \leftarrow U_q \cup \{c\}$ and continue
  end if
  $\bar{C} \leftarrow \bar{C} \cup \{\bar{c}\}$
  $\ell \leftarrow \phi(\bar{c}, F(G))$
  if $\ell =$ `support` **then**
    $E_q^+ \leftarrow E_q^+ \cup \{\bar{c}\}$
  else if $\ell =$ `contradict` **then**
    $E_q^- \leftarrow E_q^- \cup \{\bar{c}\}$
  else
    $E_q^u \leftarrow E_q^u \cup \{\bar{c}\}$
  end if
end for
$\mathcal{A}_q \leftarrow \textsc{Fill}(R_q, E_q^+)$
$FA(q) \leftarrow \mathbf{1}[\hat{y} = y \wedge |E_q^-| = 0 \wedge \exists A \in \mathcal{A}_q : f_q(A) = y]$
compute AA, CP, C-CUR, and M-CUR
return $\hat{y}$, claim labels, and all reported metrics

B.4 Metrics

The main metrics use the notation in Section 3.3. AA measures final-answer correctness over all benchmark questions. FA is assigned when the answer is correct, no contradicted claim is present, and the supported claims admit a complete contract assignment that derives the gold answer. C-CUR is the all-question rate for correct answers with at least one contradicted claim. M-CUR is the remaining correct-but-unfaithful mass, $\mathrm{AA} - \mathrm{FA} - \text{C-CUR}$. CP is pooled over all scorable claims across questions, $\mathrm{CP} = \sum_q |E_q^+| / \sum_q (|E_q^+| + |E_q^-|)$.

C Detailed Results

This section collects result views that complement the compact model-family and reasoning-level table in the main paper: subtype metrics and per-model claim profiles. Ranked by FA rather than AA, eight of the 22 models move by at least three positions.

Table A5: Results on Answer Accuracy (AA) by question subtype.

Model	Perception				Relational					Temporal		
	Pres.	Cnt.	Loc.	Fine	Dir.	Top.	Prox.	Proj.	Agg.	Turn.	Trans.	Net
Proprietary MLLMs												
GPT-4o	85.4	47.6	64.7	49.8	49.5	48.2	47.0	50.0	77.3	44.4	52.1	48.6
GPT-5.1	85.0	50.8	66.4	53.0	51.4	49.5	49.0	53.0	79.7	40.6	58.1	45.4
Claude Haiku 4.5	79.7	42.3	48.9	31.2	54.8	39.9	37.0	43.1	68.0	38.8	49.7	34.8
Claude Sonnet 4.6	84.6	52.4	68.9	43.3	51.9	55.0	54.5	60.4	82.0	44.7	55.1	52.3
Gemini 3.1 Pro	**88.6**	**58.1**	80.4	54.7	**59.6**	**57.3**	**61.5**	**65.8**	**86.0**	**49.4**	62.0	**54.0**
Open-source MLLMs												
Qwen3-VL-8B	83.3	45.9	68.5	45.3	45.7	49.5	46.0	51.0	76.2	35.6	56.3	39.4
Qwen3-VL-32B	86.2	50.2	71.5	41.7	48.1	55.0	43.0	57.4	83.1	41.9	56.6	45.7
Qwen3.5-9B	83.7	47.8	68.9	36.0	51.0	44.5	46.0	58.9	79.7	40.3	56.9	40.8
Qwen3.5-27B	85.8	52.9	71.5	42.9	57.2	53.7	47.0	61.9	84.3	44.7	56.3	49.4
Qwen2.5-VL-7B	79.3	39.7	65.1	40.9	46.2	45.9	31.0	43.6	75.6	31.6	52.1	32.5
Qwen2.5-VL-72B	80.5	42.6	62.6	40.9	54.8	48.6	40.5	50.5	82.6	30.3	53.6	35.3
InternVL3.5-8B	85.0	54.8	79.6	50.6	49.5	50.0	51.5	58.9	80.2	36.9	50.6	31.3
InternVL3.5-38B	81.7	56.2	80.9	50.6	50.0	54.1	48.5	58.4	84.9	33.8	53.6	35.3
LLaVA-OneVision-7B	80.9	36.8	56.6	47.8	42.8	45.9	38.5	40.6	73.8	34.4	51.8	33.3
GLM-4.5V	84.6	50.9	70.2	47.8	50.0	53.2	44.0	53.0	81.4	46.2	56.9	47.1
Remote sensing MLLMs												
RSThinker	75.2	48.7	66.4	37.2	35.6	32.1	27.0	28.2	71.5	24.4	40.7	27.6
GeoChat	54.1	29.4	44.7	34.8	44.2	22.0	24.5	30.7	61.0	27.2	38.0	25.9
VHM	**88.6**	34.2	51.1	47.8	34.1	35.8	20.0	26.7	57.0	26.9	28.9	29.6
GeoReason	65.2	39.9	58.9	**58.5**	45.1	51.8	32.9	44.9	63.1	39.9	50.2	42.3
GeoZero-8B	77.2	27.7	**91.5**	28.8	28.0	53.6	42.3	56.4	56.0	29.4	**62.8**	45.4
RemoteReasoner	86.8	28.8	60.5	40.1	40.2	48.1	39.8	56.1	84.6	35.8	40.0	33.2
RS-EoT-7B	61.9	18.0	64.2	29.6	34.9	23.3	4.2	34.6	68.0	13.3	13.1	10.4

C.1 Subtype Results

Question-type results cover the 12 subtypes used in the benchmark: object presence, object counting, object localization, fine-grained recognition, directional relation, topological relation, proximity, projective ordering, aggregate distribution, category turnover, semantic transition, and net change. This view supports the subtype-level analysis in Figure 6 and identifies whether a fidelity gap is mainly contradiction-bearing or missing-support.

Tables A5–A8 provide the complete subtype-level breakdowns for AA, FA, CC, and MC. The column abbreviations are Pres. (object presence), Cnt. (object counting), Loc. (object localization), Fine (fine-grained recognition), Dir. (directional relation), Top. (topological relation), Prox. (proximity), Proj. (projective ordering), Agg. (aggregate distribution), Turn. (category turnover), Trans. (semantic transition), and Net (net change). CC and MC denote the absolute C-CUR and M-CUR rates. For AA and FA, the best (second best) value in each column is shown in bold (underlined); for CC and MC, the largest (second largest) failure value is shown in red bold (red underlined).

Subtype-level readout. The subtype tables show that the AA–FA gap is structured by evidence type. Topological relation and net change are the clearest low-FA subtypes: averaged over the 22 models, their FA is 7.9% and 4.8%, respectively. Their failure mix differs, with topological relation dominated by missing support and net change containing a larger contradiction component. Object presence and localization also separate two common regimes: presence has high AA but mostly missing support, whereas localization has a larger contradiction-bearing component.

Projective ordering and category turnover have smaller average gaps, and counting has a smaller gap than topological relation at similar AA; the subtype tables thus locate the evidence requirements behind each failure type.

C.2 Model Claim Profiles

Table A10 reports a per-model, claim-weighted diagnostic: the distribution of supported, contradicted, and uncertain labels after pooling all extracted atomic claims for each model. This corpus-

Table A6: Results on Faithful Accuracy (FA) by question subtype.

Model	Perception				Relational					Temporal		
	Pres.	Cnt.	Loc.	Fine	Dir.	Top.	Prox.	Proj.	Agg.	Turn.	Trans.	Net
Proprietary MLLMs												
GPT-4o	46.7	28.8	40.4	19.4	31.7	**17.0**	32.0	30.2	51.7	28.1	35.8	6.3
GPT-5.1	54.9	29.5	37.0	17.4	27.9	10.6	23.0	40.1	**66.3**	29.4	40.4	9.8
Claude Haiku 4.5	59.3	24.1	19.1	17.8	**37.0**	12.4	25.0	40.6	54.7	22.8	39.5	2.3
Claude Sonnet 4.6	67.9	31.7	34.5	19.0	34.6	11.5	**40.5**	54.0	63.4	25.6	33.7	1.1
Gemini 3.1 Pro	52.8	35.4	37.9	18.2	26.4	9.6	30.5	**57.9**	56.4	**42.8**	**47.9**	**13.8**
Open-source MLLMs												
Qwen3-VL-8B	51.2	25.7	34.9	16.6	15.9	10.1	31.5	41.6	58.7	21.6	27.7	6.6
Qwen3-VL-32B	61.0	27.4	32.8	17.0	16.3	9.6	33.0	46.5	**66.3**	30.3	28.9	5.5
Qwen3.5-9B	60.6	28.2	40.9	18.6	24.0	9.2	24.0	32.2	61.0	18.4	31.6	5.7
Qwen3.5-27B	49.6	29.8	27.7	16.2	26.4	4.6	17.0	50.5	43.0	31.6	25.3	10.9
Qwen2.5-VL-7B	40.2	21.6	38.7	6.5	16.8	5.0	23.0	28.7	49.4	14.4	27.4	6.9
Qwen2.5-VL-72B	46.7	22.3	20.4	8.5	14.4	6.0	13.0	20.3	57.6	17.8	29.5	6.9
InternVL3.5-8B	52.8	**36.2**	56.6	19.8	29.8	11.9	38.0	36.1	53.5	30.0	22.3	4.6
InternVL3.5-38B	55.7	35.1	51.9	25.9	28.8	11.5	24.0	20.3	54.1	27.2	27.1	7.5
LLaVA-OneVision-7B	48.8	14.9	24.7	6.5	0.5	1.8	4.5	6.9	32.0	6.9	11.7	1.4
GLM-4.5V	50.4	30.6	44.3	21.1	36.5	**17.0**	29.5	41.1	**66.3**	38.8	33.4	4.9
Remote sensing MLLMs												
RSThinker	50.4	32.8	43.0	23.1	14.9	9.6	17.5	23.3	64.0	6.9	18.1	2.9
GeoChat	42.7	9.1	34.0	13.0	0.0	0.9	1.5	5.4	2.9	0.6	2.1	0.0
VHM	**70.3**	13.8	27.2	**27.9**	7.7	2.8	3.0	6.9	4.1	8.8	7.5	0.3
GeoReason	46.9	18.0	40.4	19.1	10.1	5.7	16.1	0.5	3.2	22.6	18.0	5.7
GeoZero-8B	51.5	22.8	**77.7**	26.1	25.0	7.3	15.6	2.3	1.0	7.1	21.1	0.2
RemoteReasoner	59.4	12.2	14.4	6.6	25.5	0.0	22.7	0.0	7.0	13.4	13.7	0.6
RS-EoT-7B	31.9	7.9	29.0	6.3	10.4	0.0	0.1	0.3	0.6	8.8	3.4	2.0

Table A7: Results on contradiction-bearing correct rate (CC/C-CUR) by question subtype.

Model	Perception				Relational					Temporal		
	Pres.	Cnt.	Loc.	Fine	Dir.	Top.	Prox.	Proj.	Agg.	Turn.	Trans.	Net
Proprietary MLLMs												
GPT-4o	0.0	7.6	8.9	8.1	1.9	1.4	1.5	3.0	2.3	3.1	3.6	19.8
GPT-5.1	1.2	9.2	16.2	**23.9**	12.0	16.1	2.5	5.9	4.7	7.5	10.5	25.6
Claude Haiku 4.5	0.0	7.8	26.8	12.6	11.5	7.8	5.5	2.0	2.9	13.4	7.8	23.0
Claude Sonnet 4.6	0.4	8.4	23.4	18.6	4.3	5.0	**9.5**	4.0	7.0	**17.8**	8.4	32.8
Gemini 3.1 Pro	0.4	8.1	20.0	22.7	8.7	10.1	3.5	1.0	7.0	1.2	13.6	25.9
Open-source MLLMs												
Qwen3-VL-8B	2.0	9.1	25.5	**23.9**	20.7	9.2	7.0	3.5	2.3	9.4	13.0	22.4
Qwen3-VL-32B	2.8	**11.8**	**32.8**	22.7	**21.2**	**33.5**	6.0	5.9	5.2	10.0	14.2	**34.5**
Qwen3.5-9B	0.0	6.9	18.3	11.7	8.7	16.1	4.5	6.4	2.3	9.1	10.5	19.5
Qwen3.5-27B	1.6	10.7	25.5	**23.9**	11.5	18.3	7.5	4.0	**11.0**	7.5	10.8	26.4
Qwen2.5-VL-7B	0.0	6.5	14.5	8.1	20.2	9.6	1.5	3.5	**11.0**	6.9	9.9	11.2
Qwen2.5-VL-72B	4.1	10.3	28.5	18.2	20.7	19.7	4.5	**15.8**	7.6	8.1	7.2	16.1
InternVL3.5-8B	0.0	6.4	19.1	11.7	3.8	1.4	1.0	0.0	3.5	3.4	4.8	11.8
InternVL3.5-38B	0.0	7.2	17.0	8.9	6.2	7.8	1.5	0.5	4.1	2.2	9.9	13.5
LLaVA-OneVision-7B	0.4	7.4	14.9	21.5	3.8	7.3	7.0	1.0	9.3	5.0	9.3	12.6
GLM-4.5V	0.4	7.1	16.2	11.3	3.8	15.1	1.5	5.0	3.5	5.3	13.6	30.7
Remote sensing MLLMs												
RSThinker	0.8	5.8	16.6	8.9	5.8	2.3	3.0	0.5	2.3	10.6	7.5	8.9
GeoChat	0.8	6.7	5.1	8.5	6.7	0.9	4.5	0.5	1.7	9.1	1.5	4.3
VHM	1.2	4.6	17.4	15.4	1.9	0.0	0.0	0.0	0.6	2.5	2.4	6.0
GeoReason	0.7	3.0	14.1	14.0	11.5	23.9	5.5	2.2	0.7	11.2	10.3	17.7
GeoZero-8B	0.6	4.8	11.6	2.5	2.1	6.2	0.3	1.0	4.0	15.5	**27.2**	16.3
RemoteReasoner	0.1	2.0	19.3	7.0	5.4	6.6	4.2	12.5	0.2	3.7	9.5	7.6
RS-EoT-7B	**9.9**	5.8	27.8	13.7	16.9	11.4	2.5	0.2	9.5	1.8	9.0	6.3

level view complements the question-level AA, FA, C-CUR, and M-CUR in Table 1 and underlies the coverage analysis in Appendix D.7.

Table A8: Results on missing-support correct rate (MC/M-CUR) by question subtype.

Model	Perception				Relational					Temporal		
	Pres.	Cnt.	Loc.	Fine	Dir.	Top.	Prox.	Proj.	Agg.	Turn.	Trans.	Net
Proprietary MLLMs												
GPT-4o	38.6	11.2	15.3	22.3	15.9	29.8	13.5	16.8	23.3	13.1	12.7	22.4
GPT-5.1	28.9	12.1	13.2	11.7	11.5	22.9	23.5	6.9	8.7	3.8	7.2	10.1
Claude Haiku 4.5	20.3	10.4	3.0	0.8	6.2	19.7	6.5	0.5	10.5	2.5	2.4	9.5
Claude Sonnet 4.6	16.3	12.3	11.1	5.7	13.0	38.5	4.5	2.5	11.6	1.2	13.0	18.4
Gemini 3.1 Pro	35.4	14.6	22.6	13.8	24.5	37.6	**27.5**	6.9	22.7	5.3	0.6	14.4
Open-source MLLMs												
Qwen3-VL-8B	30.1	11.1	8.1	4.9	9.1	30.3	7.5	5.9	15.1	4.7	15.7	10.3
Qwen3-VL-32B	22.4	11.0	6.0	2.0	10.6	11.9	4.0	5.0	11.6	1.6	13.6	5.7
Qwen3.5-9B	23.2	12.7	9.8	5.7	18.3	19.3	17.5	20.3	16.3	12.8	14.8	15.5
Qwen3.5-27B	34.6	12.4	18.3	2.8	19.2	30.7	22.5	7.4	30.2	5.6	20.2	12.1
Qwen2.5-VL-7B	**39.0**	11.6	11.9	26.3	9.1	31.2	6.5	11.4	15.1	10.3	14.8	14.4
Qwen2.5-VL-72B	29.7	10.0	13.6	14.2	19.7	22.9	23.0	14.4	17.4	4.4	16.9	12.4
InternVL3.5-8B	32.1	12.2	3.8	19.0	15.9	36.7	12.5	22.8	23.3	3.4	23.5	14.9
InternVL3.5-38B	26.0	13.9	11.9	15.8	14.9	34.9	23.0	37.6	26.7	4.4	16.6	14.4
LLaVA-OneVision-7B	31.7	14.5	17.0	19.8	**38.5**	36.7	27.0	32.7	32.6	**22.5**	30.7	19.3
GLM-4.5V	33.7	13.2	9.8	15.4	9.6	21.1	13.0	6.9	11.6	2.2	9.9	11.5
Remote sensing MLLMs												
RSThinker	24.0	10.1	6.8	5.3	14.9	20.2	6.5	4.5	5.2	6.9	15.1	15.8
GeoChat	10.6	13.6	5.5	13.4	37.5	20.2	18.5	24.8	56.4	17.5	**34.3**	21.6
VHM	17.1	15.8	6.4	4.5	24.5	33.0	17.0	19.8	52.3	15.6	19.0	23.3
GeoReason	17.7	**19.0**	4.4	25.3	23.5	22.3	11.4	42.2	59.1	6.1	22.0	19.0
GeoZero-8B	25.1	0.1	2.2	0.2	0.8	40.2	26.3	**53.1**	50.9	6.9	14.6	**28.9**
RemoteReasoner	27.2	14.6	**26.8**	**26.5**	9.4	**41.5**	13.0	43.6	**77.4**	18.7	16.7	25.0
RS-EoT-7B	20.1	4.2	7.3	9.7	7.6	11.9	1.5	34.1	58.0	2.7	0.7	2.0

Claim-profile readout. Table A10 separates response style from answer-critical support. FA separates evidence quality from verbosity and from claim-level support: RS-EoT-7B states by far the most claims (29.0 per response) and has the lowest FA, Gemini 3.1 Pro (8.7 claims) leads Claude Sonnet 4.6 (13.9 claims) in overall FA, and VHM reaches 59.9% supported claims and 0.78 coverage with FA of 14.8. These profiles show how each model states visual evidence, alongside the per-question FA, C-CUR, and M-CUR scores.

C.3 Design Properties of FA

FA, CP, and AvgClaims report different layers of the stated evidence and are read jointly. The slot-filling clause of Eq. (1) takes existential form: a response satisfies it whenever some assignment $A \in \textsc{Fill}(R_q, E_q^+)$ yields $f_q(A) = y_q$, so a response with many supported claims and a response with only the answer-critical supported claim receive the same slot-filling outcome, and supported claims that bind no slot neither raise nor lower FA. The contradiction clause $|E_q^-| = 0$ evaluates the stated evidence as a whole, as a reader judges a complete explanation. CP pools supported and contradicted claims across questions, so the two metrics answer different questions: CP asks how often a stated claim is correct, and FA asks whether a correct response states the answer-critical evidence without contradiction.

The per-model values (Table A9) show the resulting profile. Among the five highest-FA models, AvgClaims spans 5.2 (InternVL3.5-8B, FA 31.9) to 13.9 (Claude Sonnet 4.6, FA 33.1), with Gemini 3.1 Pro (8.7 claims, FA 35.3), GLM-4.5V (7.8, FA 33.8), and GPT-5.1 (9.5, FA 31.3) in between; at the other end, VHM (3.1 claims) has FA 14.8, LLaVA-OneVision-7B (7.2) has 12.7, and RS-EoT-7B (29.0) has 8.7. CP lies in the 71.5–80.5% band for 20 of the 22 models (GeoZero-8B: 67.1%, RS-EoT-7B: 63.7%), while FA spans 8.7–35.3%.

D Quality Control

This section reports human and automatic checks on the measurement pipeline, contradiction labels, claim extractor, verification coverage, and prompt format.

Table A9: Joint profile of response length, faithfulness, claim precision, and failure rates. Columns are recombined from Tables 1 and A10; values are percentages except AvgClaims.

Model	AvgClaims	FA	CP	C-CUR	M-CUR
Proprietary MLLMs					
GPT-4o	5.1	30.0	80.5	5.5	19.0
GPT-5.1	9.5	31.3	77.2	11.8	12.7
Claude Haiku 4.5	12.1	28.3	71.5	11.0	7.1
Claude Sonnet 4.6	13.9	33.1	73.6	12.6	11.8
Gemini 3.1 Pro	8.7	35.3	80.1	10.7	17.8
Open-Source MLLMs					
Qwen3-VL-8B	10.2	27.3	74.8	13.2	12.0
Qwen3-VL-32B	13.4	29.9	73.9	17.4	8.3
Qwen3.5-9B	8.0	28.4	76.8	10.0	15.1
Qwen3.5-27B	10.8	27.1	76.7	13.7	17.0
Qwen2.5-VL-7B	4.9	22.4	76.1	8.9	16.3
Qwen2.5-VL-72B	7.4	21.0	74.8	13.7	15.8
InternVL3.5-8B	5.2	31.9	78.9	6.1	17.2
InternVL3.5-38B	6.4	30.3	79.2	6.9	18.6
LLaVA-OneVision-7B	7.2	12.7	79.9	8.6	26.2
GLM-4.5V	7.8	33.8	78.0	10.0	12.5
Remote Sensing MLLMs					
RSThinker	7.6	24.2	78.1	6.7	11.1
GeoChat	5.9	9.2	71.5	4.5	22.1
VHM	3.1	14.8	76.7	4.8	19.6
GeoReason	7.6	17.8	73.2	10.0	21.2
GeoZero-8B	8.8	22.2	67.1	8.5	18.6
RemoteReasoner	6.0	14.4	77.7	6.8	26.7
RS-EoT-7B	29.0	8.7	63.7	9.8	11.6

D.1 Human Audit

Audit Setup. We audit responses from five representative models on a shared stratified subset of 500 questions spanning the twelve question subtypes: GPT-5.1 and Claude Sonnet 4.6 (proprietary), Qwen2.5-VL-72B (open-source), and RSThinker and GeoChat (remote sensing-specific). Each selected model is evaluated on all 500 questions, yielding 2,500 model responses. Three annotators with graduate-level training in remote sensing or computer vision conduct two audit layers on this response pool. Two primary annotators independently review each response; the third adjudicates support-rating disagreements, and contract-aware scores are their mean. The audit annotators are independent of question generation, question validation (Appendix A.4), and contract design, and in the schema-blind layer they see neither the evidence slots nor the automated labels.

Schema-blind Support Rating. Annotators first rate each response on a three-point support scale (*fully* / *partially* / *not supported*) without access to the slot schema. Pre-adjudication agreement for this support-scale rating is Cohen's $\kappa = 0.71$. After adjudication, we binarize *fully* as positive and *partially* / *not supported* as negative, then compare the result with the automated FA decision; the agreement is 89.4% with Cohen's $\kappa = 0.74$.

Pipeline Consistency. In a separate contract-aware audit, annotators rescore each stage of the automated pipeline—claim extraction, slot normalization, support/contradict/uncertain verification, slot filling, and FA/C-CUR/M-CUR assignment—under the same evidence contracts and fact layer used by the automated verifier. Per-model deltas relative to the automated scores are reported in Table 3 and remain within ± 0.8 points on FA, C-CUR, and M-CUR.

D.2 Contradiction Audit

For each audited claim labeled `contradict`, annotators verify whether the conflict is confirmed by the image and source annotation, falls outside a closed verification scope (and should be reassigned to `uncertain`), or reflects missing source annotation. Annotators confirm 95.2% of the audited contradiction labels; applying their corrections changes C-CUR by 0.4 points on the audited subset.

Table A10: Per-model distribution of verifier labels over extracted claims. Supported, Contradicted, and Uncertain are pooled over all extracted atomic claims for each model and sum to 100 per row after rounding. Coverage denotes the claim-weighted share of supported or contradicted claims, and AvgClaims denotes the average number of extracted claims per response.

Model	AvgClaims	Supp.	Contr.	Uncert.	Cov.
Proprietary MLLMs					
GPT-4o	5.1	63.8	15.4	20.8	0.79
GPT-5.1	9.5	51.9	15.3	32.8	0.67
Claude Haiku 4.5	12.1	45.4	18.0	36.6	0.63
Claude Sonnet 4.6	13.9	47.2	16.9	35.9	0.64
Gemini 3.1 Pro	8.7	52.4	13.0	34.6	0.65
Open-Source MLLMs					
Qwen3-VL-8B	10.2	50.8	17.1	32.1	0.68
Qwen3-VL-32B	13.4	44.4	15.7	39.9	0.60
Qwen3.5-9B	8.0	53.7	16.2	30.1	0.70
Qwen3.5-27B	10.8	48.9	14.9	36.2	0.64
Qwen2.5-VL-7B	4.9	64.3	20.2	15.5	0.84
Qwen2.5-VL-72B	7.4	53.5	18.0	28.5	0.71
InternVL3.5-8B	5.2	66.4	17.8	15.8	0.84
InternVL3.5-38B	6.4	62.1	16.3	21.6	0.78
LLaVA-OneVision-7B	7.2	45.3	11.4	43.3	0.57
GLM-4.5V	7.8	56.2	15.9	27.9	0.72
Remote Sensing MLLMs					
RSThinker	7.6	55.3	15.5	29.2	0.71
GeoChat	5.9	50.5	20.2	29.3	0.71
VHM	3.1	59.9	18.2	21.9	0.78
GeoReason	7.6	39.1	14.3	46.6	0.53
GeoZero-8B	8.8	41.5	20.4	38.1	0.62
RemoteReasoner	6.0	38.7	11.2	50.1	0.50
RS-EoT-7B	29.0	19.9	11.3	68.8	0.31

Table A11: Human trust study on answer-matched response pairs (%). Both responses in a pair give the correct answer to the same question.

Comparison	Pairs	Faithful preferred	Unfaithful preferred	Tie
Faithful vs. C-CUR	200	81.5	8.5	10.0
Faithful vs. M-CUR	200	69.0	17.0	14.0

D.3 Human Trust Study

The trust study asks whether the three response outcomes matter to human readers when answer accuracy scores two responses as equal.

Pair construction. From the full-benchmark responses of 18 MLLMs spanning the three model classes, we build 200 faithful–C-CUR pairs and 200 faithful–M-CUR pairs. Both responses in a pair answer the same question, and both answers map to the gold answer. Pairs are matched on the number of extracted claims and then on the token length of the visual-evidence spans. Each question and each response appears in at most one pair. Sampling is stratified to cover all three reasoning levels and to balance model families and source models across the two sides of the pairs. Of the 400 pairs, 76.8% have the same number of extracted claims, and the median absolute difference in evidence length is 6 tokens.

Protocol. Three annotators with remote sensing training who took no part in benchmark construction view the image, the question, and both responses; model identities and automated labels are hidden, and the left–right order is randomized. Each annotator independently answers: "Considering the image and question together with each response's final answer and stated visual evidence, which response is more trustworthy?" The options are Response A, Response B, and Tie. The majority vote gives the outcome of a pair, and a pair without a majority counts as a tie. Agreement is Fleiss' $\kappa = 0.63$.

Table A12: Extractor accuracy by claim group on the 500-response gold set.

Claim group	Gold claims	Precision (%)	Recall (%)	F_1
Fine-grained spatial relations	643	87.5	85.2	0.864
Exact and bounded quantities	517	91.9	90.1	0.910
Synonymous and fine-grained classes	829	89.2	87.3	0.882
Temporal descriptions	1,047	86.3	81.8	0.840

Result. Answer accuracy scores both responses in every pair as correct, yet annotators prefer the faithful response in 81.5% of faithful–C-CUR pairs and 69.0% of faithful–M-CUR pairs (Table A11). Readers thus trust verified evidence: among answers that accuracy scores as equal, they prefer the faithfully supported one.

D.4 Visual Dependence

On the 500-question subset used by the human audit, we remove the image while keeping the question and the evidence-request prompt fixed. AA falls from 33.6–57.4% with the image to 20.2–27.3% without it (Figure 7), at or below the 25–33% chance level of the answer formats, so answering requires the image.

D.5 Extractor Quality

On a 500-response subset sampled from the audited response pool, annotators independently produce gold atomic claims for each response without reference to extractor output. The frozen extractor (Qwen3-32B) achieves $F_1 = 0.91$ against these gold claims.

Table A12 breaks this comparison down for four groups of expressions that are most prone to parsing and normalization errors. Predicted and gold claims are matched one-to-one within each response, and a match requires the same claim type, time anchor, and every relevant normalized field. Spatial claims must preserve the ordered subject, relation, and object. Class claims must match the canonical category and, where stated, the fine-grained subtype, which tests the normalization of synonyms and subtype mentions. Temporal claims must keep the t_1, t_2, or pair anchor. Exact quantities must match the value, and bounded quantities must match the comparison operator and object; vague quantity expressions such as "several" are not normalized into counting claims. All four groups reach $F_1 \geq 0.84$, led by exact and bounded quantities (0.910).

D.6 Extractor Robustness

We compare two fixed claim extractors, Qwen3-32B and GPT-4.1, on responses from all 22 models to the same stratified 500-question subset. Questions, responses, verifier, and scoring rules are shared, so the comparison isolates claim extraction and normalization; Table A13 reports both scores for each model.

Across the 22 models, the mean and maximum absolute FA changes are 1.20 and 2.40 percentage points, with Spearman $\rho = 0.966$ and a maximum rank change of 6.0 positions (average ranks for ties). For M-CUR, the corresponding changes are 0.52 and 1.80 points, with $\rho = 0.988$ and a maximum rank change of 4.0 positions. Both the FA ordering and the missing-support ordering are therefore highly consistent across these two extractor choices on the matched responses. The gold-claim comparison (Appendix D.5) and end-to-end human audit (Appendix D.1) provide complementary checks against independently annotated evidence.

D.7 Verification Coverage

For each model, we report the claim-weighted share of extracted atomic claims that receive a `support` or `contradict` label, computed by pooling claims across all responses for that model: $(\sum_q |E_q^+| + \sum_q |E_q^-|)/(\sum_q |E_q^+| + \sum_q |E_q^-| + \sum_q |E_q^u|)$. Across the 22 models, this corpus-level coverage ranges from 0.31 to 0.84 (median 0.68, interquartile range 0.62–0.72); per-model values are listed in Table A10. The model with the highest FA, Gemini 3.1 Pro (FA 35.3), has below-median coverage (0.65), while two models with above-median coverage—VHM (0.78) and Qwen2.5-VL-7B (0.84)—have FA of only 14.8 and 22.4. InternVL3.5-38B and VHM share identical coverage

Table A13: Extractor comparison on 500 matched questions, all 22 models (%). Each paired cell is Qwen3-32B / GPT-4.1; AA is fixed because the model responses are unchanged. CC and MC denote C-CUR and M-CUR.

Model	AA	CP	FA	CC	MC
Proprietary MLLMs					
GPT-4o	54.0	81.1 / 83.0	29.9 / 28.3	5.2 / 6.6	18.9 / 19.1
GPT-5.1	57.4	76.8 / 78.7	27.0 / 27.4	17.2 / 15.6	13.2 / 14.4
Claude Haiku 4.5	46.6	71.8 / 76.5	25.4 / 27.1	14.8 / 13.0	6.4 / 6.5
Claude Sonnet 4.6	55.0	73.9 / 76.1	28.6 / 30.8	15.0 / 12.8	11.4 / 11.4
Gemini 3.1 Pro	64.2	80.6 / 81.0	36.0 / 34.4	13.0 / 13.8	15.2 / 16.0
Open-source MLLMs					
Qwen3-VL-8B	52.6	75.7 / 79.1	27.2 / 29.3	13.4 / 11.6	12.0 / 11.7
Qwen3-VL-32B	56.6	73.9 / 77.0	28.9 / 30.5	19.6 / 18.0	8.1 / 8.1
Qwen3.5-9B	55.2	77.9 / 79.6	28.8 / 29.4	11.2 / 9.8	15.2 / 16.0
Qwen3.5-27B	60.8	77.2 / 80.9	27.9 / 29.9	15.4 / 11.6	17.5 / 19.3
Qwen2.5-VL-7B	43.4	74.5 / 77.2	20.7 / 20.6	7.6 / 7.0	15.1 / 15.8
Qwen2.5-VL-72B	50.2	74.6 / 76.7	20.2 / 22.6	15.2 / 12.9	14.8 / 14.7
InternVL3.5-8B	55.4	80.2 / 82.8	31.8 / 32.3	6.4 / 5.6	17.2 / 17.5
InternVL3.5-38B	55.6	79.8 / 82.7	29.5 / 31.3	8.0 / 5.8	18.1 / 18.5
LLaVA-OneVision-7B	48.0	79.4 / 81.5	14.0 / 12.0	11.2 / 13.6	22.8 / 22.4
GLM-4.5V	57.8	79.6 / 83.3	34.6 / 34.5	10.4 / 9.4	12.8 / 13.9
Remote sensing MLLMs					
RSThinker	42.0	77.3 / 79.2	22.4 / 24.5	7.4 / 5.5	12.2 / 12.0
GeoChat	33.6	71.4 / 72.2	8.6 / 8.6	3.8 / 4.4	21.2 / 20.6
VHM	36.8	77.0 / 77.5	14.2 / 13.8	4.3 / 4.8	18.3 / 18.2
GeoReason	47.4	74.1 / 79.0	15.6 / 15.3	8.1 / 9.4	23.7 / 22.7
GeoZero-8B	48.9	67.3 / 71.2	21.4 / 22.7	9.6 / 8.3	17.9 / 18.0
RemoteReasoner	48.6	75.8 / 80.6	12.9 / 13.5	8.0 / 6.7	27.7 / 28.5
RS-EoT-7B	29.5	64.2 / 68.9	8.9 / 7.9	10.4 / 10.9	10.2 / 10.7

Table A14: Breakdown of verifier-labeled uncertain claims. Shares are claim-weighted over uncertain claims pooled across 18 MLLMs spanning the three model classes; OOV denotes out-of-vocabulary.

Category	Share (%)
Scene-graph entity grounding gap	24.9
OOV subject/object	18.6
Missing evidence channel	17.4
Vague / hedged / local-scope semantics	16.5
OOV predicate/attribute	9.2
Unresolved/generic referent	5.5
OOV qualifier/value	4.5
Open-scope relation/predicate	3.2

(0.78) but differ by 15.5 points in FA (30.3 vs. 14.8), an FA gap several times larger than the maximum extractor-induced FA shift reported in Table A13 ($\pm$2.4 points). FA thus separates models across the full range of coverage.

To further characterize what the uncertain bin contains, we assign each verifier-labeled uncertain claim to its dominant source of unverifiability. Table A14 reports the aggregate composition, claim-weighted across these models. Categories tied to source-annotation or vocabulary coverage account for 74.6% of uncertain claims (scene-graph entity grounding gap, missing evidence channel, and three out-of-vocabulary categories combined), reflecting situations in which $F(G)$ does not contain a verification target compatible with the asserted fact. Unmapped extraction/normalization failures are tracked separately before verifier labeling; they account for 0.4% of extracted candidate claims and are not part of the support/contradict/uncertain label distribution. Open-scope relation/predicate claims, whose predicate is in vocabulary while the queried scope is open, contribute 3.2% of uncertain claims.

D.8 Prompt Sensitivity

Table A15: Ablation study on prompt-format sensitivity using a 500-question same-sample subset. Variant A is the evidence-first baseline used in the main experiments. All variants use the same extractor and verifier. Delta columns report within-model changes relative to Variant A, and Cl. denotes the average number of extracted visual claims per response.

Model	AA	FA	CC	MC	Cl.
Variant A: Evidence-First Baseline					
GeoChat	33.6	8.6	3.8	21.2	6.1
VHM	36.8	14.2	4.3	18.3	3.7
RSThinker	42.0	22.4	7.4	12.2	11.5
Qwen2.5-VL-72B	50.2	20.2	15.2	14.8	10.2
GPT-5.1	57.4	27.0	17.2	13.2	13.4
Gemini 3.1 Pro	64.2	36.0	13.0	15.2	11.4

Model	AA	FA	CC	MC	Cl.	ΔFA	ΔCC	ΔCl.
Variant B: Answer-First								
GeoChat	33.4	8.1	4.2	21.1	6.3	-0.5	+0.4	+0.2
VHM	37.0	14.8	4.1	18.1	3.9	+0.6	-0.2	+0.2
RSThinker	42.2	21.7	7.9	12.6	11.9	-0.7	+0.5	+0.4
Qwen2.5-VL-72B	50.0	21.0	14.8	14.2	10.5	+0.8	-0.4	+0.3
GPT-5.1	57.1	26.3	17.8	13.0	13.1	-0.7	+0.6	-0.3
Gemini 3.1 Pro	64.0	36.9	12.6	14.5	11.7	+0.9	-0.4	+0.3

Model	AA	FA	CC	MC	Cl.	ΔFA	ΔCC	ΔCl.
Variant C: Open-Format								
GeoChat	32.8	7.0	4.0	21.8	5.4	-1.6	+0.2	-0.7
VHM	36.1	12.6	4.8	18.7	3.2	-1.6	+0.5	-0.5
RSThinker	41.4	20.3	7.2	13.9	10.6	-2.1	-0.2	-0.9
Qwen2.5-VL-72B	49.1	18.4	15.4	15.3	9.1	-1.8	+0.2	-1.1
GPT-5.1	56.6	24.6	17.0	15.0	12.0	-2.4	-0.2	-1.4
Gemini 3.1 Pro	63.3	33.7	13.3	16.3	10.2	-2.3	+0.3	-1.2

Prompt Variants

Variant B: Answer-first
Question: `{question}`
Provide your answer to the question, followed by the visual evidence from the image that supports your answer. Begin the answer with "Answer:" and the evidence with "Evidence:".

Variant C: Open-format
Question: `{question}`
Answer the question based on the image and explain your reasoning.

D.9 Fixed Evidence Budget

To compare FA rankings under a common requested evidence budget, we evaluate 18 MLLMs spanning the three model classes on the stratified 500-question subset. The exact-5 instruction asks for exactly five numbered atomic visual claims before the final answer. We compare these responses with the evidence-first baseline on the same questions. This control fixes the requested number of claims while leaving models to choose which visual facts to state.

Under this common instruction, FA rankings retain Spearman $\rho = 0.87$ and four of the top five models, with a mean absolute FA change of 2.8 points (Table A16). A common requested claim budget thus preserves the broad model ordering observed with freely sized evidence responses.

D.10 Answer Format

We compare multiple-choice questions with matched open-ended questions on the stratified 500-question subset using eight models spanning the three model classes. The same question program renders each open-ended question as a self-contained direct question without candidate answers.

Table A16: FA comparison under the exact-5 evidence instruction, 18 models on 500 matched questions. Absolute FA changes are percentage points relative to the evidence-first baseline.

Quantity	Result
Spearman correlation of FA rankings	0.87
Top-five overlap	4/5
Maximum absolute rank change	6
Mean absolute FA change	2.8
Maximum absolute FA change	8.1
Exact-5 format compliance (%)	95.7

Table A17: Multiple-choice / open-ended results on 500 matched questions (%). Every cell reports multiple-choice / open-ended; gap = AA − FA. The final row averages the eight model scores within each format.

Model	AA	FA	Gap	C-CUR	M-CUR
GPT-5.1	57.4 / 42.8	27.0 / 22.7	30.4 / 20.1	17.2 / 10.7	13.2 / 9.4
Gemini 3.1 Pro	64.2 / 49.8	36.0 / 30.8	28.2 / 19.0	13.0 / 8.0	15.2 / 11.0
Qwen3.5-27B	60.8 / 47.8	27.9 / 25.8	32.9 / 22.0	15.4 / 9.6	17.5 / 12.4
Qwen3-VL-32B	56.6 / 43.5	28.9 / 26.1	27.7 / 17.4	19.6 / 11.7	8.1 / 5.7
InternVL3.5-38B	55.6 / 39.0	29.5 / 24.2	26.1 / 14.8	8.0 / 4.3	18.1 / 10.5
RSThinker	42.0 / 29.4	22.4 / 17.3	19.6 / 12.1	7.4 / 4.1	12.2 / 8.0
GeoChat	33.6 / 21.8	8.6 / 7.2	25.0 / 14.6	3.8 / 2.0	21.2 / 12.6
RemoteReasoner	48.6 / 33.2	12.9 / 11.3	35.7 / 21.9	8.0 / 4.3	27.7 / 17.6
Mean	**52.4 / 38.4**	**24.2 / 20.7**	**28.2 / 17.7**	**11.6 / 6.8**	**16.6 / 10.9**

The image, gold answer, evidence contract, and evidence-request prompt remain fixed. Predefined rules map open answers into the same canonical answer space as the multiple-choice options, and both formats are evaluated on the same questions. The open-format prompt in Appendix D.8 varies the response instruction on the multiple-choice questions.

Table A17 shows that the mean gap narrows from 28.2 to 17.7 points when answer options are removed, which accounts for 75% of the drop in mean AA. Every evaluated model retains an open-ended gap, ranging from 12.1 to 22.0 points. For open-ended questions, FA, C-CUR, and M-CUR make up 53.8%, 17.8%, and 28.4% of mean AA: 46.2% of correct open-ended responses remain contradiction-bearing or missing-support. The answer format changes the size of the gap, while correct-but-unfaithful responses occur in both formats.

E Case Studies

This section presents concrete model responses under the same evaluation format: image, question, gold answer, model answer, extracted evidence claims, verification labels, filled slots, and final metric outcome.

E.1 Representative Cases

The representative set (Figures 9–11) covers one faithfully supported answer and two contradiction-bearing correct answers. Together, they show how the same answer-level outcome can receive different faithfulness labels once the response is checked against the required evidence slots.

E.2 Failure Types

The cases isolate two response-level regimes. In the counting example, all extracted visual claims are supported and collectively reconstruct the answer-critical count, so answer accuracy and faithful accuracy agree. In the existence and directional examples, the answer is still correct and the required slot can be filled, but a single contradicted claim changes the outcome to C-CUR. This distinction is why RSFaith-Bench treats evidence faithfulness as a response-level property rather than a final-answer score or an average over isolated claims.

F LIMITATIONS AND BROADER IMPACT

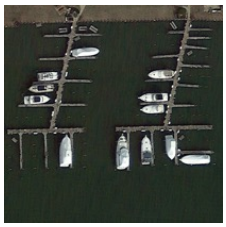
Limitations. RSFaith-Bench verifies stated evidence against the fact vocabulary and the closed scopes of its six source datasets. Claim extraction relies on a frozen language model; replacing it changes FA by at most 2.4 points in the matched-response comparison (Appendix D.6). Coverage is bounded by the source annotations: richer attributes, continuous measurements, and multi-hop spatial composition lie outside the current contracts. Scores characterize the model versions and settings in Appendix B.1; Appendix D.9 and Appendix D.10 examine changes to the evidence budget and answer format. The trust study compares blinded, answer-matched responses with the imagery available to raters.

Broader Impact. RSFaith-Bench is intended as an auditing tool for expert-facing remote sensing MLLMs, where users may rely on stated visual evidence in settings such as disaster assessment, environmental monitoring, and infrastructure inspection. By separating final-answer correctness from evidence fidelity, the benchmark can help identify models whose explanations contain contradicted or missing visual support before such systems are considered for operational use. Deployment additionally calls for task-specific human review, legal compliance, and domain validation, since remote sensing applications can affect public resources, privacy, and safety. Because remote sensing data may also be used in surveillance-sensitive contexts, releases and downstream uses of the benchmark should respect the licenses, attribution requirements, and access policies of the underlying public datasets.

Counting

Question: How many ships are visible in the image?
A. 10
B. 11
C. 12
D. 13

Answer: C



Model Answer

Evidence: The image displays two main pier structures with various ships docked at them. Scanning the left pier from top to bottom, there is one horizontally docked ship on the right side, followed by three ships docked on the left side, and one vertically docked ship at the bottom section. On the right pier, moving top to bottom, there are three ships docked on the left side of the main walkway. At the bottom section of the right pier, there are three vertically docked ships and one horizontally docked ship on the far right.
Answer: C

Evidence Contract and Verification

A. Required slot

- **S1: COUNT(ship, image, t1) = 12**
  The response must provide sufficient supported evidence to verify the total number of ships in the image.

B. Claim-by-claim verification

Claim ID	Extracted claim	Verification
C1	The image displays two main pier structures with various ships docked at them	Supported
C2	On the left pier, there is one horizontally docked ship on the right side.	Supported
C3	On the left pier, three ships are docked on the left side.	Supported
C4	On the left pier, there is one vertically docked ship at the bottom section	Supported
C5	On the right pier, there are three ships docked on the left side of the main walkway	Supported
C6	At the bottom section of the right pier, there are three vertically docked ships.	Supported
C7	At the bottom section of the right pier, one horizontally docked ship on the far right.	Supported

Slot filling for S1: 1 + 3 + 1 + 3 + 3 + 1 = 12

Metric Outcome

AA = 1 | **FA = 1** | **C-CUR = 0** | **M-CUR = 0**

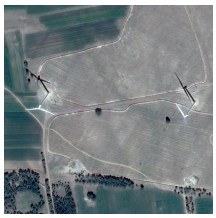
The model answer is correct, every extracted claim is supported, and the required counting slot is fully covered by the verified evidence.

Figure 9: Faithful counting case. The model gives the correct answer, and its supported count claims fill the required counting slot, yielding AA = 1 and FA = 1.

Existence

Question: Does the image contain any windmill?
A. Yes
B. No
C. Can't Judge

Answer: A



Model Answer

Evidence: The image shows an aerial view of farmland divided into multiple plots, with paved tracks intersecting in the central area. Three windmill towers are visible across the central portion of the field, each with characteristic rotor blades extending from the top. Small access roads connect the towers to the surrounding parcels.
Answer: A

Evidence Contract and Verification

A. Required slot

- **S1: EXISTENCE(windmill, image, t1) = true**
  The response must provide supported visual evidence that at least one windmill is present in the image

B. Claim-by-claim verification

Claim ID	Extracted claim	Verification
C1	windmill towers are visible across the central portion of the field.	Supported
C2	Three windmill towers are visible across the central portion of the field.	Contradicted
C3	A row of trees runs along the bottom of the image.	Uncertain

Although the response correctly implies windmill presence and thus supports **S1**, it also contains a **contradicted claim (C2).**

Metric Outcome

AA = 1 | **FA = 0** | **C-CUR = 1** | **M-CUR = 0**

The model gets the answer right, but the response is unfaithful because it includes contradicted visual evidence.

Figure 10: Contradiction-bearing existence case. The answer is correct because windmills are present, but the response overstates the number of windmills, producing a contradicted visual claim and a C-CUR outcome.

Directional

Question: In this picture, what is the position of the running track in relation to the building?
A. Below the building
B. Above the building
C. To the left of the building
D. To the right of the building

Answer: C



Model Answer

Evidence: There are a running track and a small gray-roofed building in the image. The building is near the right-center of the sports complex, while the red running track is visible toward the lower-left area. The running track is to the left of the building. The building is also close to the top-left part of the complex, near the upper fields.
Answer: C

Evidence Contract and Verification

A. Required slot

- **S1: EXISTENCE(running_track, image, t1) = true**
- **S2: EXISTENCE(building, image, t1) = true**
- **S3: REL(running_track, left_of, building, t1) = true**

B. Claim-by-claim verification

Claim ID	Extracted claim	Verification
C1	a running track in the image	Supported
C2	a small gray-roofed building in the image	Supported
C3	The building is near the right-center of the sports complex.	Supported
C4	The red running track is visible toward the lower-left area.	Supported
C5	The running track is to the left of the building	Supported
C6	The building is also close to the top-left part of the complex, near the upper fields.	Contradicted

Although **C1, C2, C5** fill the required slots **S1, S2, S3**, the response also contains a **contradicted claim (C6).**

Metric Outcome

AA = 1 | **FA = 0** | **C-CUR = 1** | **M-CUR = 0**

The model gets the answer right, but the response is unfaithful because it includes contradicted visual evidence.

Figure 11: Contradiction-bearing directional-relation case. Supported claims establish the correct left-of relation, but an additional contradicted location claim prevents the response from being counted as faithful.