Review

From Perception to Reasoning in Remote Sensing: A Survey and Outlook

Jiancheng Pan [*,‡] , Liang Yao [‡] , Zilun Zhang [‡] , Yijie Zheng , Jiahao Li , Xiao He , Wenjia Xu , Yuqian Fu , Fan Liu , Zhaojun Liu , Jianwei Yin , Xiaomeng Huang [*]

Posted Date: 26 August 2026

doi: 10.20944/preprints202608.1840.v1

Keywords: remote sensing; earth observation; vision–language models; multimodal reasoning; geospatial intelligence; reinforcement learning; tool-augmented agents

 doi:10.20944/preprints202608.1840.v1

Review

From Perception to Reasoning in Remote Sensing: A Survey and Outlook †

Jiancheng Pan [1,*,‡], Liang Yao [2,‡], Zilun Zhang [3,‡], Yijie Zheng [4], Jiahao Li [1], Xiao He [5], Wenjia Xu [6], Yuqian Fu [7], Fan Liu [2], Zhaojun Liu [8], Jianwei Yin [3], Xiaomeng Huang [1,*]

[1] Department of Earth System Science, Tsinghua University, Beijing 100084, China.
[2] College of Computer Science and Software Engineering, Hohai University, Nanjing 2111000, China.
[3] Zilun Zhang and Jianwei Yin is with Zhejiang University, Hangzhou 310058, China
[4] School of Electronic, Electrical and Communication Engineering, University of Chinese Academy of Sciences, Beijing 101408, China.
[5] Institute of Artificial Intelligence, the School of Computer Science, Wuhan University, Wuhan 430070, China.
[6] Beijing University of Posts and Telecommunications, Beijing 100190, China.
[7] KAUST, Saudi Arabia.
[8] China Academy of Space Technology, Beijing 100094, China
* Correspondence: jiancheng.pan.plus@gmail.com (J.P.); hxm@tsinghua.edu.cn (X.H.)
† This work was supported by the National Natural Science Foundation of China (42125503, 42430602).
‡ Jiancheng Pan, Liang Yao and, Zilun Zhang contributed equally.

Abstract

With the rapid growth of remote sensing data and multimodal information, foundation models for Earth observation and vision–language geospatial reasoning models have become key drivers of intelligent remote sensing analysis. However, the specific roles and applications of reasoning models in remote sensing remain insufficiently explored. This survey defines RS-Reasoning operationally as task-dependent, multi-step inference whose conclusions are supported by traceable visual, temporal, spatial, or tool-derived evidence and are evaluated beyond final-answer accuracy. We organize the literature into three non-exclusive paradigms according to the dominant mechanism by which reasoning is acquired and executed: supervised reasoning, reinforcement learning–driven reasoning, and agentic or tool-augmented reasoning. We further distinguish reasoning-specific methods from vision–language models that provide semantic alignment, generation, or grounding foundations but do not themselves demonstrate multi-step inference. Across urban analysis, disaster assessment, environmental monitoring, and spatiotemporal question answering, we identify the conditions under which reasoning may add value while separating benchmark capability from validated operational deployment. Our synthesis exposes four persistent limitations: frequent use of synthetic supervision with unevenly reported provenance, outcome-dominant evaluation, weak verification of cross-modal evidence, and fragile long-horizon tool use. We therefore outline a research agenda centered on geography-aware process supervision, multimodal constraint checking, calibrated uncertainty, expert feedback, and robustness tests across sensors, regions, and time. For convenient reference and further research, we maintain a curated collection of related resources at Awesome-RS-Reasoning-Models.

Keywords: remote sensing; earth observation; vision–language models; multimodal reasoning; geospatial intelligence; reinforcement learning; tool-augmented agents

1. Introduction

Remote sensing now combines expanding satellite archives [1–5], flexible high-resolution observations [6] from aircraft and unmanned aerial vehicles [7–9], and complementary ground-based sensors [10]. These platforms provide different spatial scales, revisit patterns, and measurement conditions. In parallel, foundation and vision–language models have expanded semantic alignment,

 doi:10.20944/preprints202608.1840.v1

open-ended interaction, and transfer across Earth observation tasks [11,12]. Together, these developments make cross-source analysis increasingly feasible, but they do not by themselves establish evidence-dependent reasoning.

Figure 1 provides an indicative view of this convergence: approximately two-thirds of the Google Scholar title records matching both "remote sensing" and "reasoning" since 2020 appeared in 2025–2026. These records measure terminology uptake rather than the number of studies satisfying our operational RS-Reasoning criteria; that gap motivates a scope-controlled synthesis of what current methods actually reason over, expose, and verify.

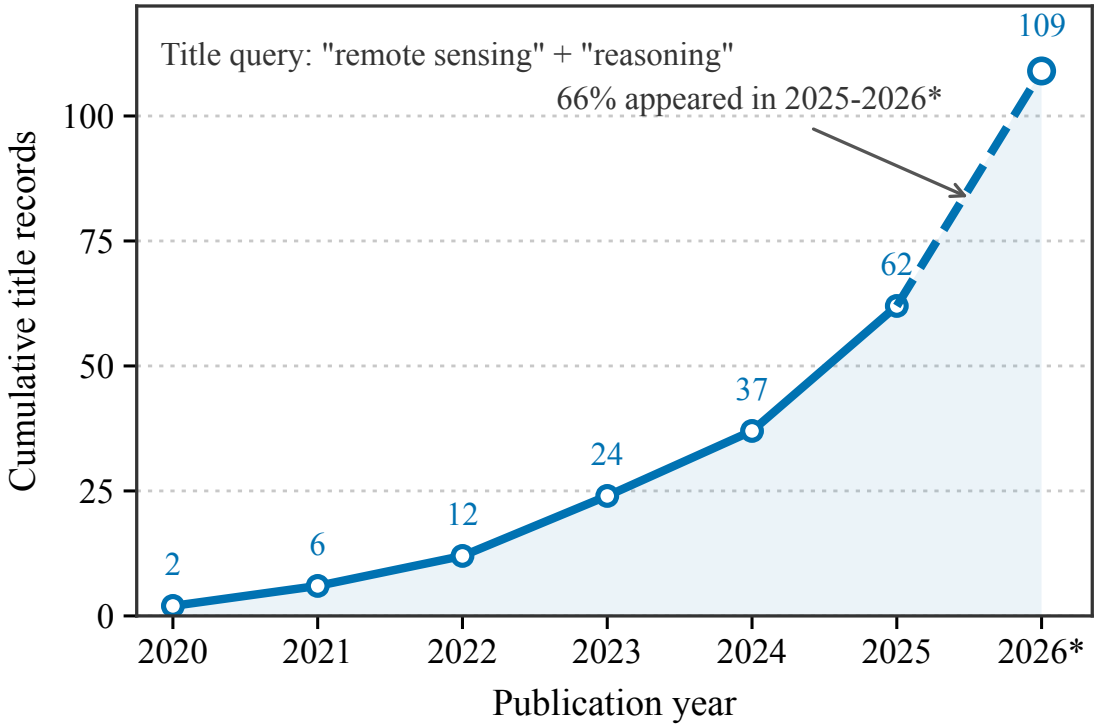


Figure 1. Approximate cumulative number of Google Scholar records whose titles contain both "remote sensing" and "reasoning," obtained by summing single-year searches from 2020 onward. Patents and citation-only records are excluded; 2026* is year-to-date through August, 2026. These records are not the survey's screened inclusion count.

Most existing remote sensing intelligence is evaluated on perception-centric tasks such as land-cover classification [13–15], object detection [16–21], and semantic segmentation [22–24]. These systems estimate what is present and where, but they do not by themselves establish how observations, spatial constraints, temporal records, or external data support a task-dependent conclusion. Urban analysis, disaster assessment, and environmental monitoring may additionally require comparison across dates and modalities, explicit geographic operations, competing-hypothesis checks, or executable GIS steps [25–28]. We therefore treat reasoning as an evidence-integrating layer over perception, not as a replacement for perception or, by itself, evidence of causality or operational decision quality.

Adjacent surveys provide three important foundations. Remote sensing language and vision–language surveys organize tasks, datasets, architectures, training strategies, and temporal capabilities [11,29–31]; broader foundation-model surveys map the Earth observation model landscape [12]; and general multimodal-reasoning surveys synthesize reasoning methods and evaluation beyond the geospatial domain [32]. Ma et al. [33] further compare domain-specific and general-purpose MLLMs and discuss spatial reasoning, reliable evaluation, and tool-augmented agents. This survey occupies the intersection of these lines: it specifies operational criteria for evidence-dependent inference, organizes methods by acquisition or execution mechanism and observable granularity, and evaluates grounding,

 doi:10.20944/preprints202608.1840.v1

geographic validity, and tool execution under remote sensing constraints. Table 1 summarizes this positioning.

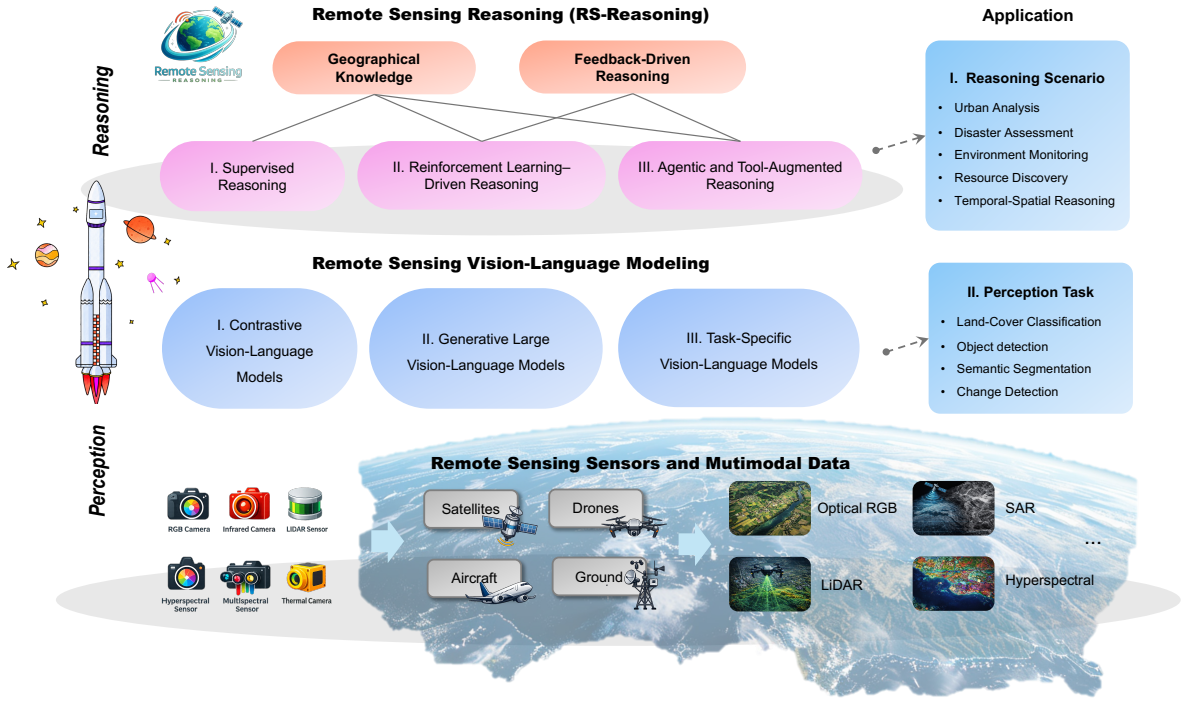


Figure 2. Conceptual scope of the survey. Multisensor Earth observation data support perception and vision–language modeling, which provide enabling representations and task interfaces. The upper layer shows three non-exclusive mechanisms through which reasoning behavior is acquired or executed: supervised, reinforcement learning–driven, and agentic/tool-augmented reasoning. Geographic knowledge and feedback are cross-cutting signals rather than separate stages. The right panel distinguishes perception tasks from application settings in which relational, temporal, or executable analysis may be required. Vertical placement denotes analytical scope, not a chronological or maturity progression, and vision–language modeling alone is not evidence of reasoning.

We use the operational criteria formalized in Section II to distinguish reasoning-specific systems from enabling vision–language foundations. We then analyze the former through three non-exclusive paradigms, classified by the dominant mechanism through which behavior is acquired or executed: supervised reasoning, reward-based post-training that optimizes task outcomes or process proxies, and agentic/tool-augmented reasoning that plans or executes external operations. A method may span more than one paradigm; these labels identify its principal mechanism rather than a chronological stage. A second axis records where evidence, predictions, or actions become observable, allowing mechanism and output granularity to be compared separately.

Our contributions are fourfold:

- We provide an operational definition and explicit inclusion criteria for RS-Reasoning, distinguishing multi-step, evidence-dependent inference from perception, semantic alignment, grounding alone, and fluent rationale generation.
- We separate general-purpose multimodal datasets from task-oriented benchmarks, preserve heterogeneous supervision units without invalid aggregation, and organize reasoning methods through a two-axis taxonomy of acquisition/execution mechanism and observable evidence/output granularity.
- We provide an evidence-calibrated synthesis of urban analysis, disaster assessment, environmental monitoring, and spatiotemporal question answering, separating demonstrated benchmark or workflow capabilities from unsupported causal, operational, or autonomous-decision claims.

 doi:10.20944/preprints202608.1840.v1

- We introduce a capability-to-evidence evaluation matrix and a falsifiable research roadmap that pair claimed advances with required evidence, diagnostic interventions, invalid shortcuts, and predefined success criteria.

Table 1. Positioning relative to adjacent surveys. Only the axes needed to distinguish scope, organizing principle, and reasoning-process coverage are retained.

Survey	Scope	Primary organizing axis	Reasoning-process coverage
Bashmal et al. [29]	Language integration in RS	Tasks, datasets, and open challenges	Covers language-conditioned interpretation broadly; reasoning mechanisms and executable workflows are not central.
Li et al. [11]	RS vision–language models	Model architectures and downstream tasks	Reasoning appears as an enabling capability; no reasoning-specific taxonomy or agent-execution analysis.
Weng et al. [30]	RS vision–language models	Contrastive learning, instruction tuning, and text-conditioned generation	Discusses model capability and explanation-driven reliability, while reasoning mechanisms and executable agents are not the primary organizing axes.
Liu et al. [31]	Spatiotemporal RS VLMs	Temporal tasks, datasets, methods, and metrics	Centers temporal interpretation, but not a general taxonomy of reasoning acquisition or tool-augmented execution.
Xiao et al. [12]	RS and EO foundation models	Broad foundation-model landscape	Comprehensive task coverage, but reasoning and execution are not the organizing axes.
Wang et al. [32]	General multimodal reasoning	Reasoning methods and evaluation	Reasoning is central, but sensor, scale, and geospatial constraints are outside its scope.
Ma et al. [33]	RS MLLM image understanding	Domain-specific versus general-purpose models	Diagnoses grounding, spatial reasoning, and reliable evaluation, and discusses tool-augmented agents; reasoning acquisition and execution are not the organizing axes.
This survey	**RS reasoning**	**Acquisition/execution mechanism × observable evidence/output granularity**	**Operational definition, two-axis taxonomy, evidence-calibrated applications, capability-to-evidence evaluation, and verifier/execution requirements.**

We classify a study as reasoning-specific only when (1) at least one conclusion depends on combining multiple evidence items or inference steps; (2) intermediate claims or actions can be linked to observations, constraints, or tool outputs; and (3) evaluation probes more than final-answer correctness, such as grounding, process validity, calibration, or trajectory execution. Studies that provide alignment, generation, detection, or grounding without satisfying these criteria are treated as enabling foundations.

The remainder of this paper is organized as follows. Section II formalizes the scope and operational criteria of RS-Reasoning, and Section III motivates evidence-integrating analysis under remote sensing constraints. Section IV reviews multimodal datasets and benchmarks, remote sensing vision–language models, and post-training foundations. Section V introduces the unified two-axis taxonomy and examines the three complementary mechanisms. Section VI evaluates four application settings through demonstrated evidence and unsupported decision boundaries. Section VII analyzes data, model design, scientific validity, training objectives and verifiers, reproducibility, efficiency, and governance. Section VIII concludes with a falsifiable research roadmap.

2. Definition and Scope: What Is Remote Sensing Reasoning?

We define remote sensing reasoning (RS-Reasoning) as task-dependent, multi-step inference in which a system combines Earth observation evidence with geographic, temporal, or domain constraints to derive a conclusion that is not available from a single perception output. The conclusion must be accompanied by an auditable support structure, such as grounded regions, cited observations, intermediate spatial relations, executed tool calls, or calibrated uncertainty. Cross-modal semantic integration [34–36] is therefore necessary but not sufficient: aligning imagery and text becomes reasoning only when the aligned evidence is used in a query-dependent inference process.

 doi:10.20944/preprints202608.1840.v1

This definition separates three levels of capability. **Perception** extracts observable elements such as classes, objects, regions, or changes. **Understanding** aggregates those elements into scene-level semantics. **Reasoning** additionally tests relations or hypotheses under explicit constraints and produces conclusions whose validity depends on the evidence chain. The relevant capabilities include relational and compositional inference, temporal ordering, cross-modal evidence reconciliation, and causal hypothesis assessment. We use *causal hypothesis assessment* deliberately: observational imagery alone rarely identifies a causal effect, so causal claims should be conditioned on assumptions and corroborated with temporal records, physical models, interventions, or external data. Typical tasks include evidence-grounded visual question answering, multi-temporal event interpretation, cross-image hypothesis testing, and joint visual–language–GIS analysis [27,28,37].

Consider a post-flood observation of a river basin. A perception model delineates water and damaged vegetation; an understanding model labels the scene as a flood. A reasoning system compares pre- and post-event imagery, queries a digital elevation model and river-network topology, and checks whether discharge records support a dam-release hypothesis against alternatives such as extreme rainfall or levee failure. It then reports downstream exposure by elevation and road accessibility together with the evidence and uncertainty for each inference. The added value is not a longer textual explanation, but a conclusion that can be inspected, challenged, and recomputed when evidence or assumptions change.

Operationally, we classify a system as reasoning-specific when: (1) at least one conclusion depends on combining multiple evidence items or inference steps; (2) intermediate claims can be linked to observations, constraints, or tool outputs; and (3) the system is evaluated for more than final-answer correctness, for example through grounding, process validity, counterfactual consistency, calibration, or trajectory execution. A fluent chain of thought without these properties is not sufficient evidence of reasoning, and a deterministic GIS workflow may still implement reasoning if its task-dependent choices and intermediate conclusions satisfy these criteria.

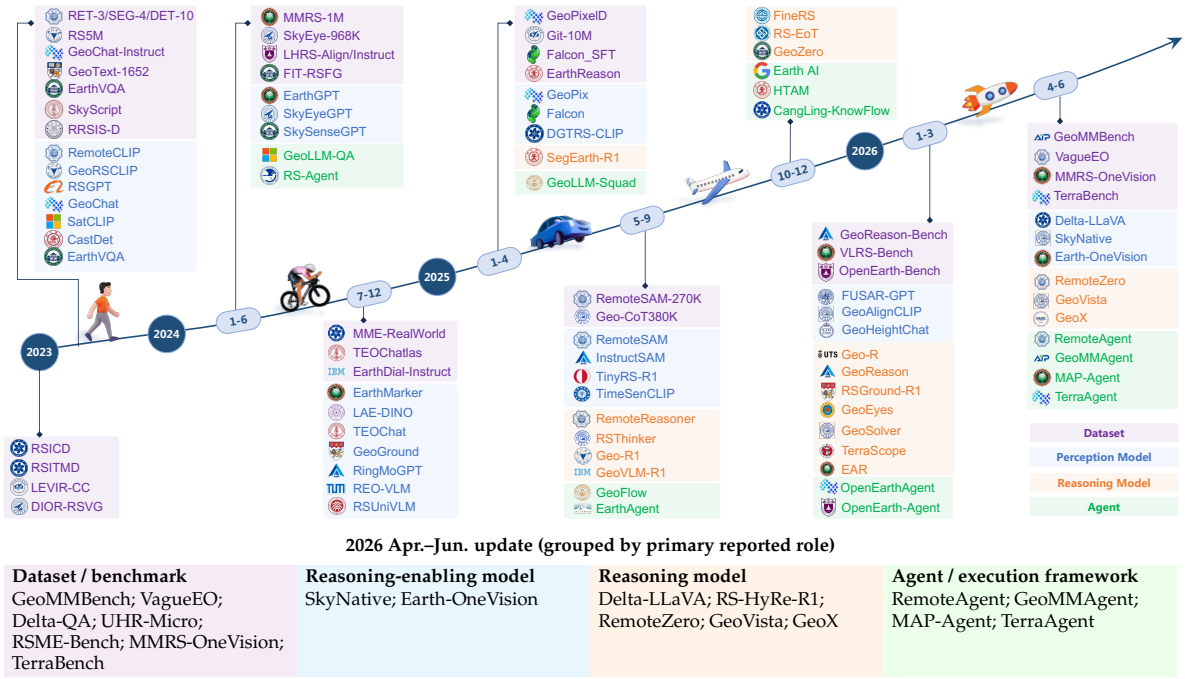


Figure 3. Timeline of representative remote sensing vision–language and reasoning work (2023–June 2026). The lower strip adds papers first posted from April through June 2026 and lists companion resources under the paper that introduced them [38–48]. The figure traces the expansion from perception-oriented methods toward reasoning-specific and agentic paradigms, with datasets, models, and benchmarks grouped by their primary role. Placement on the timeline indicates publication chronology rather than a claim that later systems necessarily provide stronger or more faithful reasoning.

 doi:10.20944/preprints202608.1840.v1

3. Motivation: Why Do We Need Reasoning in Remote Sensing?

3.1. Why Perception Is Not Enough

The remote sensing community has achieved substantial progress in perception-centric tasks such as land-cover classification [13–15], object detection [18,20,21,49–52], and semantic segmentation [22–24]. These capabilities provide the indispensable observational layer of Earth intelligence: they identify what is present, locate relevant objects and regions, and measure visible changes. Yet many scientific and operational questions begin only after these outputs have been produced. A land-cover map does not determine whether an urban expansion pattern is compatible with transport accessibility; a change mask does not identify whether a damaged region reflects flooding, fire, structural collapse, or recovery; and a spectral anomaly does not establish which environmental process produced it. Such questions require models to relate observations across locations, times, modalities, and external sources rather than interpret each prediction in isolation.

Reasoning is therefore not a replacement for perception, but a mechanism for converting perception outputs into task-dependent conclusions. It connects detected evidence with contextual knowledge, compares competing interpretations, applies spatial or temporal constraints, and identifies what additional information is needed before a conclusion can be supported. This distinction is especially important for application-driven remote sensing, where analysts commonly ask relational, explanatory, and decision-support questions: which communities are exposed after a hazard, which observations support a proposed change mechanism, or which candidate regions satisfy several geographic conditions simultaneously. The importance of RS-Reasoning lies in enabling this transition from isolated predictions to evidence-integrating analysis.

3.2. Why Remote Sensing Reasoning Is Distinctive

Remote sensing reasoning is not simply generic visual reasoning applied to overhead imagery. Earth observations are indirect, scale-dependent measurements acquired by heterogeneous sensors under changing atmospheric, seasonal, and geometric conditions. Relevant evidence may be distributed across optical, SAR, hyperspectral, thermal, elevation, and time-series data, and the same phenomenon may appear differently across spatial resolutions or acquisition times. Consequently, a plausible conclusion often depends on reconciling incomplete or conflicting modalities, preserving spatial correspondence across scales, and distinguishing genuine temporal change from sensor or acquisition effects. These properties make evidence integration itself a central difficulty rather than a secondary modeling choice.

Geographic structure further distinguishes RS-Reasoning from conventional image understanding. Many conclusions depend on topology, distance, direction, containment, upstream–downstream relations, administrative boundaries, or physical constraints that are only partly visible in the imagery. An environmental analyst may need to combine spectral evidence with river-network connectivity and land-use records; an urban analyst may combine land cover with roads, population, and points of interest; and a disaster analyst may compare multi-temporal observations with elevation and infrastructure data [25–27]. Reasoning provides a framework for composing these heterogeneous sources while keeping their assumptions and dependencies explicit. It also clarifies the limits of inference: observational remote sensing can support and challenge hypotheses, but causal or policy conclusions generally require corroborating records, domain models, or expert judgment.

3.3. Why Reasoning Matters Now

Recent advances make this transition technically plausible. Remote sensing vision–language models provide increasingly strong semantic alignment, open-ended interaction, and spatial grounding [11,12]. Reasoning-oriented post-training and reinforcement learning begin to optimize multi-step task behavior, evidence seeking, and process consistency rather than language generation alone [53,54]. In parallel, tool-augmented agents can connect foundation models [55] with GIS operators, remote sensing algorithms, spatial databases, and domain-specific products, allowing natural-language re-

 doi:10.20944/preprints202608.1840.v1

quests to be translated into inspectable analytical workflows [27,28,37]. The convergence of these developments creates an opportunity to move from models that describe remote sensing images toward systems that organize evidence and execute geospatial analyses.

The resulting value should be understood in scientific and operational terms. A useful reasoning system can make a conclusion easier to inspect by exposing supporting observations, spatial relations, intermediate products, and tool outputs; it can make an analysis easier to revise when evidence or assumptions change; and it can reduce the gap between an analyst's question and a reproducible sequence of geospatial operations. These benefits are particularly relevant to disaster response, environmental monitoring, infrastructure analysis, and policy support, where unsupported but fluent answers may carry substantial consequences. The objective is therefore not to generate longer rationales or to claim autonomous decision making, but to develop remote sensing systems that help experts formulate hypotheses, integrate heterogeneous evidence, test constraints, and produce more transparent and reproducible analyses. This emerging role motivates a dedicated review of how RS-Reasoning is learned, executed, evaluated, and connected to real Earth observation applications.

4. Foundation: Remote Sensing Vision-Language Modeling and General Vision-Language Reasoning Models

4.1. Remote Sensing Multimodal Datasets and Benchmarks

Multimodal resources play two related but distinct roles in remote sensing vision–language research. General-purpose datasets supply supervision for representation learning, pre-training, and transfer, whereas task-oriented benchmarks define the inputs, outputs, and evaluation protocols used to test particular capabilities. We discuss these roles separately because corpus scale alone does not determine benchmark quality, and strong performance on a narrowly defined benchmark does not necessarily imply a broadly useful training resource.

General-Purpose Multimodal Datasets

As summarized in Table 2, the development of general-purpose datasets has progressed along three main dimensions: scale, semantic richness, and sensor diversity. BigEarthNet-MM [56] established a large paired Sentinel-1/Sentinel-2 archive with multilabel land-cover supervision, providing a multisensor basis for classification and retrieval. RS5M [57] and SkyScript [58] subsequently expanded image–text pre-training to millions of pairs by combining large remote sensing image collections with natural-language or map-derived descriptions. More recent datasets place greater emphasis on annotation depth. ChatEarthNet [59] provides detailed descriptions with manual verification, LuoJiaHOG [60] organizes geo-aware captions hierarchically, and VRSBench [61] combines captions, object references, and question–answer pairs to support several vision–language tasks within one resource.

The modality coverage is also broadening beyond conventional optical image–text pairs. Landsat30-AU [62] targets long-term Landsat understanding, GAIA [63] combines multiple sensors and spatial scales, and SAR-TEXT [64] and SARLANG-1M [65] introduce language supervision for SAR imagery. Text2Earth [66], in contrast, uses resolution-aware descriptions to support text-conditioned image generation [67]. MMRS-OneVision extends this trend to roughly 34 million question–answer instructions spanning six sensor modalities, cross-sensor fusion, nine task categories, and 25 subtasks [47]. Its scale largely reflects collection, format conversion, and synthesis over public resources, however, and should not be read as 34 million independent observations. These resources expose an important trade-off: automatic caption generation enables scale, whereas manual verification, object-level grounding, and expert annotation improve semantic reliability but are considerably more expensive. Accordingly, the sample counts in Table 2 are not directly comparable, since they refer to different units and may contain overlapping image sources. Dataset size and nominal modality coverage should therefore be interpreted as properties of the supervision resource rather than direct evidence of cross-modal reasoning ability.

 doi:10.20944/preprints202608.1840.v1

Table 2. Representative general-purpose multimodal remote sensing datasets. Years and venues follow the cited paper versions; arXiv denotes that no formal publication venue was confirmed. Heterogeneous units and possible source overlap prevent direct scale comparisons.

Name	Year	Venue	#Samples	Modality	Annotation / Supervision	Main Usage
BigEarthNet-MM [56]	2021	IEEE GRSM	590K pairs	S1/S2	Multi-label LULC labels	Classification, retrieval
RS5M [57]	2024	IEEE TGRS	5.0M pairs	Image/Text	English descriptions	Pre-training, retrieval, ZSC
SkyScript [58]	2024	AAAI	2.6M pairs	Image/Text	OSM-grounded semantic text	General VLM pre-training
ChatEarthNet [59]	2025	ESSD	163K/10K pairs	S2/Text	GPT-generated captions with manual verification	Geo-foundation pre-training
LuoJiaHOG [60]	2025	ISPRS JPRS	-	Image/Text	Geo-aware hierarchical captions	Image-text retrieval
VRSBench [61]	2024	NeurIPS	30K img; 123K QA	Image/Text	Captions, object refs, QA pairs	Captioning, grounding, VQA
BigEarthNet.txt [68]	2026	arXiv	464K pairs; 9.6M texts	S1/S2/Text	Captions, VQA, referring expressions	Instruction tuning, multi-task VLM
Landsat30-AU [62]	2026	AAAI	96K pairs; 18K VQA	Landsat/Text	Captions + human-verified VQA	Long-term VL understanding
GAIA [63]	2026	IEEE GRSM	201K pairs	Multi-sensor/Text	Curated text + synthetic captions	Classification, retrieval, captioning
MMRS-OneVision [47]	2026	arXiv	∼34M QA	Six sensors + fusion/Text	Collected, converted, and synthesized instructions	Multi-sensor and cross-sensor instruction tuning
SAR-TEXT [64]	2025	arXiv	>130K pairs	SAR/Text	Generated SAR descriptions	Retrieval, captioning, VQA
SARLANG-1M [65]	2026	IEEE TGRS	>1.0M pairs	SAR/Text	Fine-grained captions + QA	SAR VLM pre-training
Text2Earth (Git-10M) [66]	2025	IEEE GRSM	10.5M pairs	Image/Text	Resolution-aware captions	Text-to-image generation

Task-Oriented Multimodal Benchmarks

Task-oriented benchmarks, summarized in Table 3, convert broad multimodal capability claims into explicit evaluation problems. Early resources such as RSICap in RSGPT [69] and FIT-RS in SkySenseGPT [36] emphasize captioning and instruction following. Subsequent benchmarks move toward settings in which an answer must depend on temporal, spectral, spatial, or operational evidence. SECOND-CC [70] evaluates bi-temporal change captioning, while DisasterM3 [71] and DisasterInsight [72] assess multimodal disaster understanding. HM-Bench [73] targets hyperspectral question answering, and VLRS-Bench [74] and OmniEarth [75] broaden evaluation toward complex reasoning and robustness. GroundSet [76] introduces vector-grounded spatial understanding, whereas NeSy-Route [77], RSRCC [78], and GTPBD-MM [79] test constrained planning, regional change comprehension, and multimodal parcel extraction, respectively.

The April–June 2026 resources add three more evaluation targets. Delta-QA couples bi- and tri-temporal questions with point, box, text, and change-mask supervision, while UHR-Micro and RSME-Bench test whether models use small native-resolution evidence and remain visually anchored under degradation or misleading text [40,43,45]. VagueEO evaluates vague-intent interpretation and capability-aware routing, GeoMMBench contributes 1,053 expert-written questions across remote sensing, photogrammetry, GIS, and GNSS, and TerraBench exposes executable Earth-system tasks with reference tool steps and tolerance-aware numerical outcomes [38,39,48]. Their scales count different units and must not be pooled.

Unlike a pre-training corpus, the value of a benchmark depends less on its raw size than on the validity of its task formulation and evaluation protocol. Reliable assessment requires geographic and sensor-disjoint splits, controls for source overlap and instruction leakage, and metrics that separate perception errors from reasoning errors. For grounding and planning tasks, final-answer accuracy should be complemented by localization quality, constraint satisfaction, intermediate-decision validity, and robustness to missing or conflicting modalities. The progression in Table 3 therefore reflects more than an increase in task variety: it marks a transition from evaluating language generation about remote sensing imagery to evaluating whether model outputs are supported by spatially and physically relevant evidence.

 doi:10.20944/preprints202608.1840.v1

Table 3. Task-oriented multimodal remote sensing benchmarks grouped by evaluation focus. Scale is reported in the native units of each source and is not directly comparable across rows. Formal venues are used when confirmed; otherwise arXiv is shown.

Benchmark	Year	Venue	Scale	Inputs	Evaluated capability
Language supervision and generation					
RSGPT (RSICap) [69]	2025	ISPRS JPRS	2.6K captions	Image + text	Captioning
SkySenseGPT (FIT-RS) [36]	2024	arXiv	1.8M instructions	Image + text	Instruction following
Temporal and disaster understanding					
SECOND-CC [70]	2025	IEEE JSTARS	6K pairs; 30K captions	Bi-temporal + text	Change captioning
DisasterM3 [71]	2025	NeurIPS	27K image pairs; 123K instructions	Multi-sensor + text	Disaster assessment
DisasterInsight [72]	2026	arXiv	112K instances	Image + text	Grounded disaster analysis
RSRCC [78]	2026	arXiv	126K questions	Bi-temporal + text	Change question answering
Delta-QA [40]	2026	arXiv	180,876 QA	Bi-/tri-temporal + prompts	Change QA and dense masks
General multimodal reasoning					
HM-Bench [73]	2026	arXiv	19K QA pairs	HSI + text	Hyperspectral reasoning
VLRS-Bench [74]	2026	arXiv	2K QA pairs	Image + text	Complex reasoning
OmniEarth [75]	2026	arXiv	9K images; 44K instructions	Multi-source + text	Perception, reasoning, and robustness
RSME-Bench [45]	2026	arXiv	291 degradation images; 2,648 misleading-text QA	Image + text	Visual-reliance stress tests
Spatial grounding and planning					
GroundSet [76]	2026	arXiv	510K images; 3.8M objects	Image + vector + text	Spatial grounding
NeSy-Route [77]	2026	arXiv	11K samples	Image + mask + text	Constrained route planning
GTPBD-MM [79]	2026	arXiv	-	Image + text + DEM	Parcel and boundary extraction
Evidence acquisition and agentic execution					
VagueEO [39]	2026	arXiv	5K train; 1K test pairs	Image + vague text	Intent interpretation and tool routing
GeoMMBench [38]	2026	CVPR	1,053 MCQs	Multi-sensor images/maps + text	Expert geoscience multimodal QA
UHR-Micro [43]	2026	arXiv	1,212 images; 11,253 instructions	UHR image + text	Micro-evidence grounding and reasoning
TerraBench [48]	2026	arXiv	403 tasks; ∼24.5K steps	EO + grids + GIS + simulation	Executable workflow and numeric grounding

Table 4. Summary of Vision-Language Models in Remote Sensing. Task Type abbreviations: **ZSC** (Zero-Shot Classification), **ITR** (Image-Text Retrieval), **GeoLoc** (Geolocation), **RepL** (Representation Learning), **IC** (Image Captioning), **VQA** (Visual Question Answering), **GC** (Grounded Captioning), **CD** (Change Detection), **Gen-VL** (General Vision-Language Tasks), **OVD** (Open-Vocabulary Detection), **VG** (Visual Grounding), **RES** (Referring Expression Segmentation).

Category	**Model**	**Modality**	**Task Type**
Contrastive VLMs	RemoteCLIP [80], GeoRSCLIP [57]	RGB, Text	ZSC, ITR
	DGTRS-CLIP [81], PriorCLIP [82,83]	RGB, Text	ZSC, ITR
	SatCLIP [84]	RGB, Geo-coordinates	GeoLoc, RepL
	TimeSenCLIP [85]	Time-Series, Text	RepL
Generative Large VLMs	RSGPT [69], SkyEyeGPT [86], LHRS-Bot [87]	RGB, Text	IC, VQA
	GeoChat [35], EarthMarker [88], SkySenseGPT [36]	RGB, Prompts	VG, GC, VQA
	EarthGPT [34], TEOChat [89], RingMoGPT [90]	Optical, SAR, IR, Temporal	CD, IC, VQA
	RSUniVLM [91], Falcon [92]	Multi-Modal	Gen-VL
	SkyNative [45], Earth-OneVision [47]	Multi-sensor, Temporal	Gen-VL, CD, VG, RES
	Delta-LLaVA [40]	Bi-/tri-temporal	CD, VQA, RES
Task-Specific VLMs	CastDet [93], Cross-View OVD [94]	Aerial/Ground RGB	OVD
	LAE-DINO [20], OpenRSD [95], LLaMA-Unidetector [96], FASE [97]	RGB	OVD
	GeoGround [98], RSVG-ZeroOV [99]	RGB	VG
	RemoteSAM [100], UniGeoSeg [101], GeoMag [102], InstructSAM [103]	RGB	RES
	EarthMind [104]	Optical, SAR	RES

4.2. Remote Sensing Vision-Language Models

Contrastive Vision-Language Models

The adaptation of CLIP [105] to remote sensing (RS) aims to leverage its zero-shot transfer capabilities but faces inherent challenges posed by data scarcity and domain divergence. Recent literature primarily addresses these bottlenecks through data scaling, semantic refinement, and the

 doi:10.20944/preprints202608.1840.v1

integration of RS-specific priors. To mitigate data scarcity, methods such as RemoteCLIP [80] and GeoRSCLIP [57] establish large-scale pre-training benchmarks (e.g., RS5M) by unifying heterogeneous annotations and leveraging VLM-generated descriptions, thereby bridging the visual-linguistic gap. Complementing data-driven approaches, DGTRS-CLIP [81] enhances semantic granularity via a dual-branch framework that aligns visual features with multi-grained captions. Furthermore, recent works extend the contrastive paradigm beyond standard RGB-text pairs by incorporating intrinsic RS characteristics [106]; for instance, SatCLIP [84] encodes global geolocation embeddings, while TimeSenCLIP [85] exploits temporal and spectral consistency in satellite time-series to reduce reliance on textual supervision.

Generative Large Vision-Language Models

Whereas CLIP-based models primarily support representation alignment, large vision–language models (LVLMs) add open-ended generation and instruction-conditioned interaction. Early remote sensing adaptations, including RSGPT [69], SkyEyeGPT [86], and LHRS-Bot [87], use domain-specific instruction data and alignment modules to improve scene description. GeoChat [35], EarthMarker [88], and SkySenseGPT [36] further introduce visual prompting and region-level grounding for finer spatial interaction. Multi-sensor and temporal coverage is represented by EarthGPT [34], which accepts optical, synthetic aperture radar (SAR), and infrared inputs, and by TEOChat [89] and RingMoGPT [90], which address time-series interpretation and change analysis. RSUniVLM [91] and Falcon [92] broaden multi-granularity and general-purpose interpretation. Recent work extends the foundation in different directions. SkyNative maps native image patches into a shared autoregressive backbone and uses degradation and misleading-text tests to probe visual reliance; Earth-OneVision unifies six sensor modalities and cross-sensor fusion through multilevel feature injection and tokenized spatial outputs [45,47]. Delta-LLaVA is more reasoning-specific: it combines adjacent-frame difference features, change-guided attention, and a mask decoder under supervised bi- and tri-temporal QA, exposing both textual conclusions and dense change evidence [40]. These advances establish stronger prerequisites for reasoning, but generative fluency, region grounding, or temporal task performance should not be treated as evidence of faithful multi-step reasoning unless the intermediate evidence and process are also evaluated.

Task-Specific Vision-Language Models

Beyond general representation alignment, a subset of RSVLM research focuses on dense prediction that can physically ground later reasoning steps. In open-vocabulary object detection (OVD), CastDet [93] and Cross-View OVD [94] transfer knowledge from frozen CLIP teachers or ground-view representations to aerial imagery. LAE-DINO [20] and OpenRSD [95] introduce dynamic vocabulary construction and open-prompt frameworks, while LLaMA-Unidetector [96] and FASE [97] separate localization from recognition or incorporate scene context. For visual grounding, GeoGround [98] supports horizontal boxes, oriented boxes, and masks, whereas RSVG-ZeroOV [99] explores training-free diffusion-based refinement. For referring-expression segmentation, RemoteSAM [100] and Uni-GeoSeg [101] provide unified perception baselines; GeoMag [102], InstructSAM [103], and Earth-Mind [104] address resolution adaptation, training-free segmentation, and optical–SAR segmentation, respectively. These methods contribute localization evidence, but dense prediction alone remains a perception or grounding capability unless it participates in an evaluated inference chain.

Collectively, these models provide essential ingredients for RS-Reasoning: semantic alignment, open-ended generation, and spatial grounding. We nevertheless treat them as *reasoning-enabling* rather than automatically *reasoning-demonstrating*. Open-vocabulary recognition, captioning, visual question answering, or chain-of-thought generation does not by itself establish multi-step inference. Under the criteria in Section II, a reasoning-specific claim additionally requires evidence-dependent intermediate decisions and evaluation of grounding, process validity, calibration, or execution. This distinction prevents improvements in linguistic fluency from being conflated with improvements in geospatial reasoning.

 doi:10.20944/preprints202608.1840.v1

4.3. Post-Training Foundations for Vision-Language Reasoning

Scope and Training Routes

Reasoning-oriented post-training usually starts from a pretrained vision–language policy, with supervised fine-tuning (SFT) supplying domain knowledge, output conventions, or an initial reasoning scaffold. The classical alignment pipeline combines SFT, a learned preference model, and online policy optimization such as proximal policy optimization (PPO) [107,108]. Current multimodal systems, however, need not follow this entire sequence. Two routes should be distinguished: *offline preference optimization*, which learns from a fixed set of ranked responses, and *online reinforcement learning*, which samples new trajectories and scores them during training. This distinction is important because the presence of a policy-shaped loss does not by itself imply online exploration or verifiable reasoning.

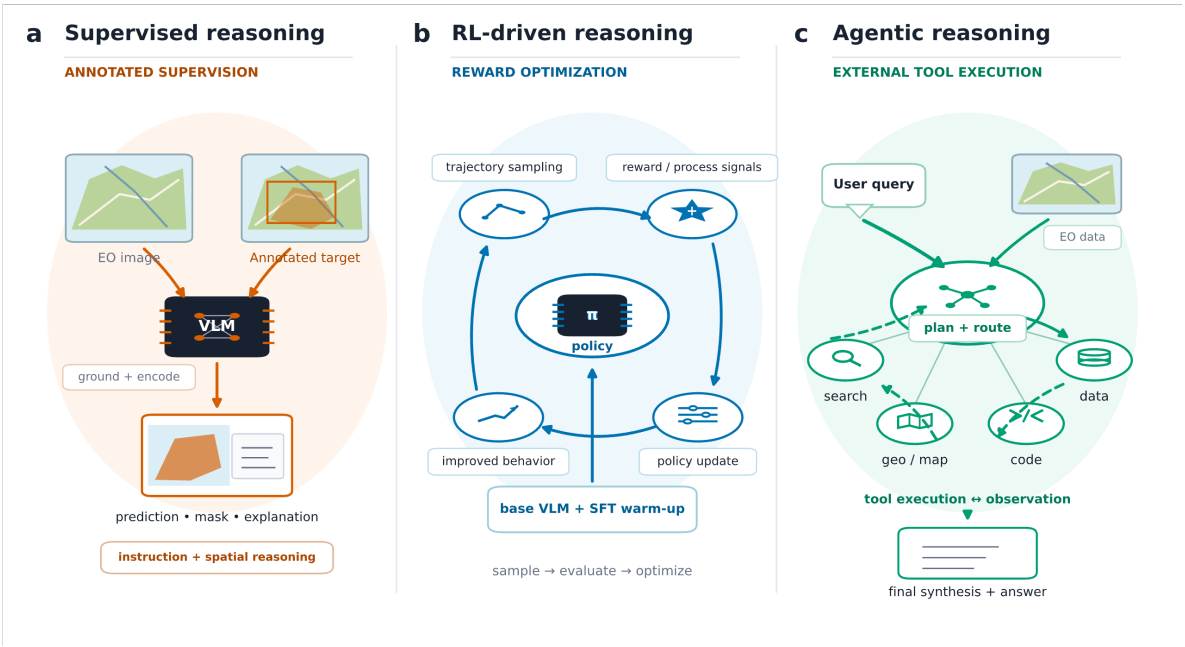


Figure 4. Representative reasoning paradigms for Earth observation. **a**, Supervised reasoning learns visual grounding and spatial inference from annotated remote-sensing data to generate predictions, segmentation masks, or textual explanations. **b**, Reinforcement-learning-driven reasoning iteratively samples reasoning trajectories, evaluates reward or process signals, and updates the policy (e.g., through GRPO or related RL algorithms) following supervised fine-tuning. **c**, Agentic or tool-augmented reasoning decomposes a user query, selects and executes external tools—including search, geospatial processing, code execution, and data access—and incorporates observations through an iterative planning–feedback loop before synthesizing the final answer. Solid arrows indicate forward information flow, whereas dashed arrows denote iterative optimization or observation–planning feedback. The three paradigms are organized according to their dominant reasoning mechanism rather than model maturity.

Offline Preference Optimization

Direct Preference Optimization (DPO) [109] is a representative offline method. For a prompt x, preferred response y_w, rejected response y_l, and reference policy π_{ref}, define $s_\theta(x,y) = \beta \log[\pi_\theta(y \mid x)/\pi_{\text{ref}}(y \mid x)]$. DPO minimizes

$$\mathcal{L}_{\text{DPO}} = -\mathbb{E}_{(x,y_w,y_l)\sim\mathcal{D}_{\text{pref}}} \log \sigma(s_\theta(x,y_w) - s_\theta(x,y_l)), \tag{1}$$

which, under the Bradley–Terry preference model, corresponds to a KL-regularized reward objective without an explicit reward model or on-policy rollout. DPO is therefore useful when pairwise judgments of answer quality, response style, or routing behavior are available, but it should be treated as a neighboring alignment technique rather than evidence that a model has learned to seek and verify new geospatial evidence online.

 doi:10.20944/preprints202608.1840.v1

Online RL with Verifiable Rewards

Online RL instead repeatedly samples candidate outputs, evaluates them, and updates the policy. PPO uses a learned value function as a variance-reducing baseline, whereas Group Relative Policy Optimization (GRPO) [110] replaces the critic with relative rewards among responses to the same query. For G sampled trajectories and task reward R_i, a useful remote-sensing abstraction and the core group-relative advantage are

$$R_i = \lambda_a R_{\text{ans}} + \lambda_f R_{\text{fmt}} + \lambda_g R_{\text{grd}} + \lambda_p R_{\text{proc}} + \lambda_t R_{\text{tool}}, \qquad \widehat{A}_i = \frac{R_i - \overline{R}}{s_R + \epsilon}, \tag{2}$$

where the active terms depend on the task, $\overline{R}$ and s_R are the group mean and standard deviation, and ϵ prevents numerical instability. The resulting loop is compact: *sample G trajectories → compute task rewards → normalize within the group → apply a clipped policy update with reference-policy regularization → resample*. Under outcome supervision, the same sequence-level advantage is normally broadcast to every token; group normalization therefore reduces the need for a critic but does not by itself solve step-level credit assignment. General multimodal methods illustrate complementary reward interfaces: Visual-RFT [111] uses verifiable visual-task rewards, while R1-VL [112] adds dense rule-based rewards for key-step accuracy and reasoning validity.

where the active terms depend on the task, $\overline{R}$ and s_R are the group mean and standard deviation, and ϵ prevents numerical instability. The resulting loop is compact: *sample G trajectories → compute task rewards → normalize within the group → apply a clipped policy update with reference-policy regularization → resample*. Under outcome supervision, the same sequence-level advantage is normally broadcast to every token; group normalization therefore reduces the need for a critic but does not by itself solve step-level credit assignment. General multimodal methods illustrate complementary reward interfaces: Visual-RFT [111] uses verifiable visual-task rewards, while R1-VL [112] adds dense rule-based rewards for key-step accuracy and reasoning validity. Concrete remote-sensing methods and their verifier scopes are compared in Section 5.2; Section 7.4 examines how these signals should be audited.

5. Taxonomy: Representative Paradigms of Remote Sensing Reasoning Models

We use one taxonomy with two orthogonal axes. Axis A records the dominant mechanism responsible for reasoning behavior: annotated demonstrations for supervised methods, reward-based policy updates for RL-driven methods, and task decomposition plus external execution for agentic methods. Axis B records where the behavior is made observable: text or reasoning trajectories, object/region grounding, pixel/dense prediction, or executable workflows. Terms such as "grounding-enhanced" and "multi-granularity" therefore describe positions on Axis B or cross-cutting design choices; they are not additional model families competing with the three paradigms. Hybrid systems receive multiple descriptors but are placed in Table 5 by their primary reported contribution. This two-axis rule makes RemoteReasoner and RemoteAgent comparable in granularity while still distinguishing policy optimization from external tool execution, and it avoids implying a universal maturity progression among the paradigms.

 doi:10.20944/preprints202608.1840.v1

Table 5. Unified two-axis taxonomy of representative remote sensing reasoning methods. Axis A identifies the dominant mechanism by which reasoning is acquired or executed; Axis B records the granularity at which evidence, predictions, or actions are expressed.

Axis A: dominant reasoning mechanism	Reasoning signal / execution pattern	Axis B: output or execution granularity	Representative methods
Supervised reasoning	Annotated answers, masks, grounding traces, or structured demonstrations	Text / QA; object or region grounding; pixel or dense prediction; multi-temporal explanation	EarthVL [113], SegEarth-R1 [114], SegEarth-R2 [115], TerraScope [26], Delta-LLaVA [40], GeoChrono [116]
RL-driven reasoning	Answer-, process-, or verifier-based rewards optimize a reasoning policy	Textual trajectory; object or region output; pixel or dense output; multi-granularity interface	GeoReason [53], GeoVLM-R1 [117], GeoZero [118], Geo-R1 [119], TinyRS-R1 [120], RSThinker [121], GeoSolver [54], RS-EoT [122], RemoteReasoner [123], RemoteAgent [39], RS-HyRe-R1 [41], RemoteZero [42], GeoVista [44], GeoX [46]
Agentic / tool-augmented reasoning	Task decomposition, retrieval, tool invocation, feedback, and executable planning	Executable workflow; tool-call sequence and arguments; multimodal evidence synthesis; final decision artifact	Earth AI [27], Earth-Agent [28], OpenEarthAgent [37], GeoMMAgent [38], EarthAgent [124], MAP-Agent [43], TerraAgent [48]

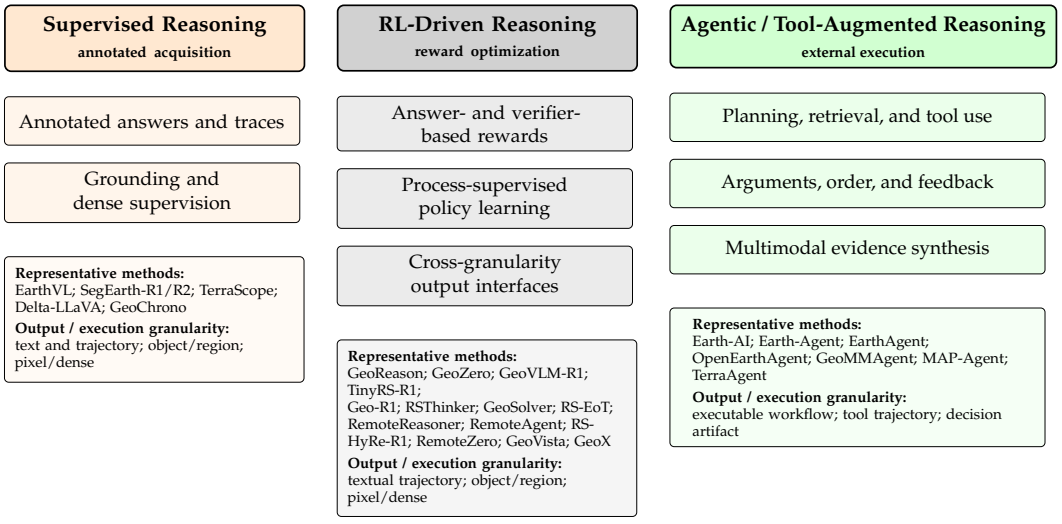


Figure 5. Unified two-axis view of remote sensing reasoning. The three columns encode the dominant acquisition or execution mechanism (Axis A), while the shared horizontal annotation records output or execution granularity (Axis B). The columns are analytical coordinates, not a maturity ranking or a mandatory progression.

5.1. Supervised Reasoning

Supervised reasoning is the most direct mechanism considered here: models learn to produce answers, masks, or structured outputs from annotated examples, instruction–response pairs, or pixel-grounded traces. The supervision can induce reasoning-like behavior without directly optimizing process validity. EarthVL [113], for example, connects semantic segmentation with visual question answering through object-level land-cover semantics. Under the operational definition in Section II, the contribution of such supervision should be assessed through the dependence of conclusions on grounded intermediate evidence, not solely through the presence of an explanatory output.

Subsequent studies extend supervised reasoning toward more explicit grounding and finer-grained spatial understanding. *SegEarth-R1* [114] introduces geospatial pixel reasoning, where the model must interpret implicit spatial language and directly output the mask of the target region, thereby expanding reasoning from textual answering to dense spatial prediction in Figure 6a. *SegEarth-R2* [115] generalizes this setting by emphasizing comprehensive language-guided segmentation under hierarchical granularity, target multiplicity, implicit intent, and linguistic variability, making the supervision signal substantially closer to real geospatial communication. In a broader Earth observation setting, *TerraScope* [26] advances supervised reasoning through pixel-grounded, modality-flexible,

 doi:10.20944/preprints202608.1840.v1

and multi-temporal reasoning, supported by large-scale reasoning traces with embedded masks. By jointly handling optical imagery, SAR, and temporal sequences, this approach expands supervised reasoning from static scene understanding to more realistic, multi-source, and change-aware analysis. Delta-LLaVA adds an explicit temporal mechanism: adjacent-frame feature differences guide attention, and a mask decoder ties textual change answers to dense bi- or tri-temporal evidence [40]. Its current evidence is nevertheless confined to benchmarks derived from semantic change labels rather than long-horizon or causal change analysis. GeoChrono extends supervised temporal reasoning to 4–19-frame histories through per-location temporal trajectories and prompt-guided token compression [116]. Its evidence is exact-match QA over NAIP-derived RGB sequences rather than verification of intermediate temporal claims or multisensor transfer. Overall, supervised reasoning methods establish the perceptual and benchmark foundations of the field. Their major strength lies in strong grounding and controllable training signals; however, their reasoning behavior remains largely imitation-based and is therefore constrained by annotation coverage, rationale quality, and the diversity of supervised tasks.

5.2. Reinforcement Learning–Driven Reasoning

Outcome- and Process-Aware Reward Optimization

Building on the post-training distinctions in Section 4.3, Table 5 compares RL-driven methods by reward observability and credit granularity rather than by optimizer name. At the outcome- and task-aware end, RSThinker applies canonical task metrics after rationale-based SFT, whereas GeoVLM-R1 combines format constraints with output-specific rewards for heterogeneous Earth observation tasks [117,121]. Geo-R1 and TinyRS-R1 show how output-level rewards can be specialized to few-shot geospatial referring (Figure 6b) and compact-model training, respectively [119,120]. Such signals can rank complete responses across output types, but an answer, overlap, or semantic-similarity score does not independently validate the intermediate geographic claims. Complementary methods place the proxy closer to the trajectory: GeoZero uses an answer-modulated heuristic thinking score without predefined rationale supervision while retaining instruction SFT; GeoReason checks conclusion stability under option permutation; and GeoSolver learns token-level process scores for tree-structured policy optimization [53,54,118]. These methods thus span endpoint ranking, controlled consistency testing, and learned within-trajectory feedback; none of these proxy signals is by itself a ground-truth verifier of the complete reasoning process.

Evidence- and Execution-Aware Optimization

A second distinction concerns where evidence and executable actions become observable. RS-EoT combines evidence-seeking trace supervision with progressive grounding and VQA reinforcement learning to train an ask–ground–answer pattern rather than a one-shot response [122]. At broader output granularity, RemoteReasoner uses a shared text- and box-generating policy, with region- and pixel-level products obtained through output transformations and external segmentation. RemoteAgent instead applies GRPO to internally handled sparse tasks and routes out-of-capability dense requests to specialist tools at inference [39,123]. These systems occupy cross-cutting positions on the survey's two axes: reward-based policy optimization describes the dominant acquisition mechanism, whereas textual trajectories, grounded predictions, derived dense outputs, and tool calls describe where behavior becomes observable. Benchmark gains or successful routing do not by themselves demonstrate faithful evidence use or robust transfer; those claims require checks of intermediate statements, tool arguments and returned values, and performance beyond the tasks, sensors, and reward functions seen during training.

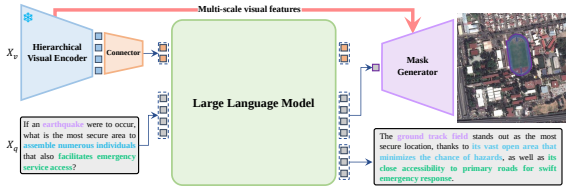
Grounded and Executable Verifier Scopes

Recent methods make the verifier boundary especially visible. RemoteZero scores a candidate box b through crop–query compatibility and an area penalty,

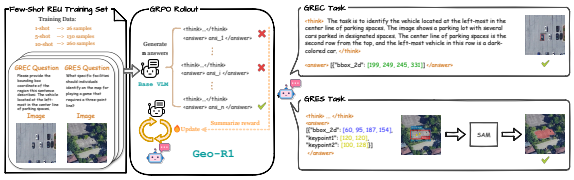
 doi:10.20944/preprints202608.1840.v1

$$r(b, Q) = s(T(I, b, \alpha), Q) - \lambda \max\left(0, \frac{|b|}{|I|} - \tau\right), \tag{3}$$

where $T(I, b, \alpha)$ crops image I around b at expansion scale α, $|b|/|I|$ is the relative box area, and τ is the tolerated area threshold [42]. This replaces target-coordinate supervision with an observable region score, but it validates the candidate region rather than the intermediate natural-language chain. GeoVista combines supervised tool-interleaved active-perception traces with GRPO for grounding and counting using format, accuracy, IoU, and state-machine plan signals [44]. GeoX uses executable geometric programs and an open-vocabulary segmenter to reward proposer–solver self-play over mask-computable spatial tasks [46], whereas RS-HyRe-R1 combines spatial-activation, perception-correctness, and visual–semantic-path rewards across grounding, open-vocabulary detection, and VQA [41]. Together, these methods expose region endpoints, crop calls and evidence-state updates, or program executions—stronger observables than an unconstrained rationale—but their verifier scopes remain bounded by crop semantics, state-machine proxies, segmenter-derived masks, or handcrafted reward components.

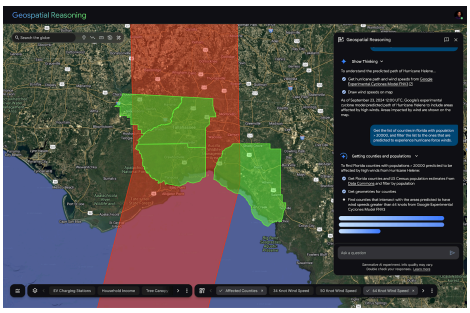


(a) SegEarth-R1 [114]: language-guided pixel reasoning with implicit spatial grounding.

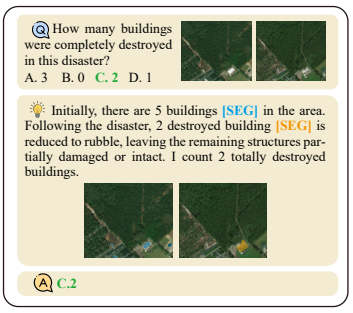


(b) Geo-R1 [119]: RL-driven reasoning for few-shot geospatial referring.

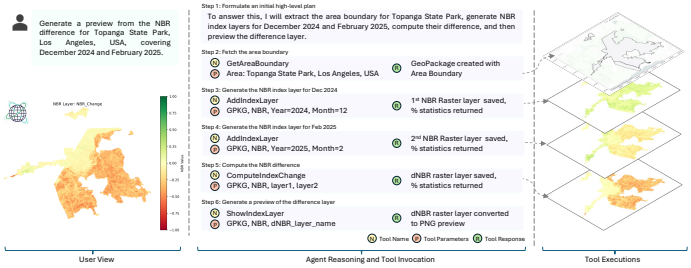
Figure 6. Representative RS reasoning models illustrating the supervised (SegEarth-R1) and RL-driven (Geo-R1) paradigms. Agentic paradigm examples (Earth AI, OpenEarthAgent) appear in the application case studies (Figure 7).



(a) Earth AI [27]: hurricane-wind and county-population intersection.



(b) TerraScope [26]: pixel-grounded destroyed-building counting.



(c) OpenEarthAgent [37]: executable NBR differencing.

Figure 7. Selected crisis-response and environmental reasoning examples. They are paper demonstrations or benchmark instances, not evidence of operational deployment, calibrated risk, or autonomous decision authority.

 doi:10.20944/preprints202608.1840.v1

5.3. Agentic and Tool-Augmented Reasoning

Agentic RS-Reasoning is best compared as an executable workflow rather than as a longer textual chain. At step t, a system may transform a request q into a plan state p_t, bind the next step to a tool u_t and arguments θ_t, receive an observation o_t, and then revise the plan or stop:

$$q \longrightarrow p_t \longrightarrow (u_t, \theta_t) \longrightarrow o_t \longrightarrow \{p_{t+1}, \text{stop}\}.$$

This abstraction is a comparison axis, not a claim that every agent implements every stage. Its value is to identify which decisions and state transitions—planning, tool choice, arguments, execution, observation use, and termination—are externally observable and therefore open to verification.

Representative systems impose orchestration structure at different levels. *Earth AI* combines a Gemini-powered reasoning agent with specialized foundation models, geospatial data sources, and tools across imagery, population, and environmental domains [27]. *GeoMMAgent* adopts a training-free plan–execute–evaluate pipeline that routes multimodal questions to retrieval, perception, and reasoning modules, followed by self-evaluation and selective re-execution [38]. *EarthAgent* takes a dependency-constrained approach: its hierarchical task abstraction mechanism derives a three-layer workflow from an expert-defined task-dependency graph to constrain inter-layer ordering [124]. The supporting evidence is complementary rather than directly comparable. Earth AI measures targeted fact-finding and rubric-based crisis-response cases, GeoMMAgent evaluates modular question answering on GeoMMBench, whereas EarthAgent evaluates tool-plan quality on GeoPlan-bench independently of actual tool execution. These results support domain-structured orchestration, but do not establish equivalent levels of end-to-end autonomy or operational reliability.

MAP-Agent represents a narrower, training-free form of visual evidence acquisition. It separates query-guided region discovery, native-resolution local inspection, and global–local evidence synthesis; SAM2 is used only to convert boxes to masks for segmentation evaluation [43]. This trajectory makes selected regions inspectable but does not constitute a general GIS tool environment. GeoMMAgent and MAP-Agent should likewise not be cited as policy-optimization methods: their reported contribution is inference-time orchestration, and their evaluations emphasize final task scores rather than learned long-horizon recovery.

Other systems focus more directly on executable tool binding. *Earth-Agent* instantiates an inference-time ReAct-style EO controller over 104 predefined MCP tools spanning index calculation, inversion, perception, analysis, and statistics; Earth-Bench evaluates multiple off-the-shelf LLM backbones using both final outcomes and reference tool trajectories [28]. *OpenEarthAgent*, by contrast, fine-tunes a Qwen3-4B controller with supervised next-tool prediction over 14,538 replay-validated structured trajectories. A unified registry and execution controller expose perceptual, GIS, spectral, and raster operations, validate arguments, execute calls, cache intermediate results, and return observations to the working context [37]. This is supervised trajectory learning rather than reinforcement learning, and deterministic replay validation should not be interpreted as deterministic or error-free agent planning.

TerraAgent broadens executable orchestration to heterogeneous Earth-system evidence through a ReAct scaffold over 77 sub-tools for EO data, environmental grids, GIS, forecasting, simulation, visualization, and document-grounded verification [48]. It materializes intermediate scientific artifacts and exposes tool arguments, order, observations, and structured outputs. TerraAgent is an execution framework rather than a separately trained policy; its evidence comes from TerraBench's process metrics and tolerance-aware numerical scores, which show that an apparently valid tool sequence need not yield a correct numerical conclusion.

Agentic orchestration need not involve executable GIS tools. As a mechanism-boundary case, *VRA* uses a training-free think–critique–act loop to coordinate off-the-shelf vision–language models, a reasoning model, and trajectory memory [125] through follow-up visual queries [126]. Its reported gains concern accuracy on a 500-question VRSBench subset at substantially higher inference cost, rather than verified tool execution or general robustness. Evaluation resources should likewise be distinguished from agent architectures. *ThinkGeo* is a domain-specific agentic benchmark, not a trained

 doi:10.20944/preprints202608.1840.v1

agent: its latest version contains 486 tasks, 1,778 expert-verified ReAct steps, and 14 executable tools, and evaluates instruction following, tool selection, argument correctness, trajectory execution, and final-answer correctness under step-wise and end-to-end protocols [127]. Together, these distinctions imply that workflow completion alone is insufficient evidence of reliable geospatial reasoning. Claims should also be scoped by execution success, observation-grounded revision, failure recovery, geographic and sensor transfer, tool cost and latency, and comparison with expert-designed GIS pipelines.

Viewed together, the three paradigms describe complementary mechanism layers rather than a strictly chronological progression. Grounded supervision couples language with visual and spatial evidence [26,40,113–115]; RL makes selected reasoning behaviors an optimization target [39,41,42,44,46,53,54,117–123,128]; and agentic systems expose task decomposition and execution through geospatial tools or active inspection [27,28,37,38,43,48,124]. None of these mechanisms guarantees faithfulness on its own. Grounding can be incomplete, a reward can be misspecified, and a correct tool can be invoked with invalid arguments. A credible hybrid system should therefore preserve traceability across the complete chain from observation and intermediate claim to verifier, tool output, and final conclusion.

6. Application: Case Studies of Remote Sensing Reasoning

Remote sensing reasoning has the potential to transform Earth observation data into decision-relevant evidence when the target is not directly observable from pixels alone and must instead be inferred from spatial context, temporal evolution, multimodal observations, or domain knowledge. The evidence base is nevertheless still dominated by retrospective benchmarks and curated demonstrations. Accordingly, this section treats urban analysis, disaster assessment, environmental monitoring, and spatiotemporal question answering as representative capability settings, not as proof that current models are ready to make autonomous operational decisions.

Table 6. Evidence-calibrated map of four RS reasoning settings. "Demonstrated evidence" records tasks or workflows actually evaluated in the cited work, whereas "decision boundary" identifies conclusions that remain unsupported.

Domain	Task requirement	Demonstrated evidence	Decision boundary
Urban & Social Analysis	Combine land-cover semantics, object relations, and explicit GIS conditions to screen candidate areas for expert review.	EarthVL evaluates segmentation-guided relational VQA in a city-planning setting; OpenEarthAgent demonstrates urban-domain POI, distance, and area operations in executable GIS tool chains [37,113].	Current evidence supports candidate screening, not transportation-behavior inference, planning eligibility, or policy benefit.
Disaster Assessment	Localize, count, and characterize damage from pre- and post-event observations while retaining spatial evidence.	TerraScope evaluates destroyed-building counts and rates and damaged-road-area estimation with pixel masks; DisasterM3 and DisasterInsight add multi-hazard, damage-level, and structured-report tasks [26,71,72].	These benchmarks do not establish causal attribution, resource-allocation priorities, or reliability in live emergency operations.
Environmental Monitoring	Combine spectral indices, temporal differences, and spatial statistics into a reproducible analytical workflow.	OpenEarthAgent executes NDVI, NBR, and NDBI workflows; Earth-Agent covers quantitative EO tools; TerraAgent combines EO, environmental grids, GIS, forecasts, and simulators [28,37,48].	Measurement and spatial association are supported; pollution-source attribution and policy conclusions remain hypotheses requiring external evidence.
Spatiotemporal QA	Link dated observations, retain frame-specific evidence, and reconstruct change histories or event order.	TEOChat evaluates temporal EO tasks, TerraScope and Delta-LLaVA ground multi-temporal change in masks, and GeoChrono/ChronoBench tests long-term histories and construction ordering [26,40,89,116].	Correct temporal order does not establish causation, permanence, or counterfactual explanations.

6.1. Urban and Social-Space Analysis

Urban reasoning is most defensible when framed as candidate screening rather than as direct inference of social behavior or planning suitability. A perception model first supplies land-cover masks or object inventories; the reasoning layer must then combine explicit relations and measurements, such as adjacency, distance, area, density, or POI context, while preserving the evidence used for each filter. *EarthVL* provides the most direct model-level evidence in this setting: its city-planning-oriented data and evaluations connect land-cover segmentation with relational multiple-choice and open-ended VQA [113]. *SegEarth-R1/R2* and *RemoteReasoner* contribute useful implicit-language

 doi:10.20944/preprints202608.1840.v1

localization and multi-granularity outputs, but their benchmarks do not evaluate an end-to-end planning workflow [114,115,123].

Executable GIS operations add a different kind of evidence. *OpenEarthAgent* includes urban-domain trajectories that call POI, distance, area, and related spatial tools, making the derived quantities and tool arguments inspectable [37]. The supported claim is therefore that current systems can translate some analyst-defined spatial criteria into reproducible screening outputs with provenance. They have not yet shown that inferred accessibility, transportation behavior, planning eligibility, or policy impact agrees with human planners or improves real decisions; those claims require strong GIS baselines, blinded expert review, and deployment-centered evaluation.

6.2. Disaster Assessment

Disaster reasoning should be separated into three evidence levels: change localization, damage characterization, and response-oriented decision support. The first two now have direct benchmark evidence. In Figure 7b, *TerraScope* compares pre- and post-event observations and grounds destroyed-building counting in pixel masks; its benchmark also includes destruction-rate and damaged-road-area estimation [26]. *DisasterM3* broadens the evidence base across hazards, sensors, structural damage, relational questions, and long-form reports, while *DisasterInsight* evaluates building function, damage level, disaster type, counting, and structured reporting [71,72]. These resources justify claims about measured benchmark capabilities, not about the correctness of downstream emergency decisions.

The Earth AI case in Figure 7a intersects predicted hurricane-force winds with county geometry and population data, but remains a targeted demonstration rather than an operational impact study [27]. *RS-EoT* improves evidence seeking on grounding and VQA benchmarks without validating disaster severity or response reliability [122]. Causal explanation or resource priority additionally requires calibrated uncertainty, independent sensor or field confirmation, and comparison with expert emergency workflows.

RemoteAgent additionally demonstrates single-step routing of vague damage-assessment requests to a specialist change-detection tool on xBD [39]. This supports intent-to-tool translation, but output quality inherits the specialist tool and the experiment does not evaluate emergency prioritization, multi-step recovery, or live operations.

6.3. Environmental Monitoring

For environmental monitoring, the strongest current evidence concerns measurement and spatial association rather than causal source attribution [129,130]. For examples, this capability is particularly important for marine and climate protection [131,132], where large-scale observations are needed to track environmental change and ecosystem degradation. A defensible workflow is *index computation* → *temporal differencing* → *spatial intersection or statistics* → *provenance-preserving artifact*. *OpenEarthAgent* directly instantiates this pattern with NDVI, NBR, and NDBI operations plus raster and GIS tools; Figure 7c, for example, shows an NBR difference workflow in which the area boundary, dated index layers, returned statistics, and preview artifact remain visible [37]. *Earth-Agent* similarly expands executable EO analysis across index calculation, inversion, spectral products, and quantitative spatiotemporal operations [28].

This evidence supports reproducible derived layers, spatial associations, and auditable tool traces. It does not show that an agent can identify why an anomaly occurred. Pollution-source attribution may additionally require hydrological flow, weather, discharge records, field samples, and alternatives such as agricultural runoff or natural blooms; vegetation-stress attribution may require climate and management records. Until those variables and competing hypotheses are evaluated, the appropriate output is a ranked, uncertainty-aware set of hypotheses linked to source data, with abstention when evidence is insufficient, rather than a single policy-grade causal conclusion.

TerraBench broadens this setting to executable workflows that combine EO imagery, gridded environmental variables, GIS operations, forecasts, and deterministic simulators while retaining numerical outputs and supporting artifacts [48]. The reported gap between process-oriented tool-use scores and

 doi:10.20944/preprints202608.1840.v1

tolerance-aware numerical scores is an important boundary: a plausible sequence of scientific tools is not sufficient evidence that the resulting environmental conclusion is numerically correct.

6.4. Spatiotemporal Narratives and Question Answering

Spatiotemporal reasoning is not a single capability. Temporal recognition asks what differs between dated observations; grounded change analysis additionally localizes the evidence; long-term reasoning must retain a history across many acquisitions and answer ordering questions. *TEOChat* establishes a temporal vision–language baseline across building change, damage assessment, semantic change detection, and temporal scene classification [89]. *TerraScope* adds frame-specific pixel grounding for multi-temporal change analysis [26]. Delta-LLaVA couples adjacent-frame feature differencing with supervised bi- and tri-temporal QA and dense change masks, providing a direct link between a textual change answer and localized evidence [40]. Its evaluation does not extend beyond three frames or establish why a change occurred. Most directly, *GeoChrono* and its ChronoBench evaluate land-cover perception, temporal recognition, long-term memory, and spatiotemporal reasoning across 12 subtasks and 17,689 validated QA pairs, including reconstruction of land-cover histories and object-construction ordering [116].

These results support a narrower and testable claim than the previous generic event-narrative framing: models can be evaluated on whether they recover dated changes and ordering relations from image sequences. A correct order does not by itself establish why an event occurred, whether a change is permanent, or what counterfactual intervention would alter it. Credible temporal answers should expose the acquisition date and localized evidence for each step, distinguish "not observed" from "did not occur," and reserve causal language for cases supported by external process knowledge.

Across these case studies, RS-Reasoning offers three testable hypotheses: it should improve performance on relational and evidence-composition queries, make conclusions more auditable through grounded evidence or execution traces, and reduce the effort required to translate analyst intent into reproducible geospatial workflows. These benefits should not be inferred from fluent rationales alone. They require prospective comparison with strong perception-plus-GIS baselines, blinded expert assessment, calibration analysis, and measurement of failure recovery, latency, and human workload. In high-stakes settings, the appropriate role of current systems is decision support with explicit provenance and abstention, not autonomous decision authority.

7. Challenge: Open Problems of Remote Sensing Reasoning

Remote sensing reasoning remains an early and methodologically heterogeneous research area. Reported benchmark gains coexist with unresolved limitations in data construction, model design, evaluation, and optimization. These limitations are amplified by ultra-high spatial resolution, geographic heterogeneity, incomplete modalities, temporal inconsistency, and the requirement for scientifically defensible inference. Progress therefore depends not only on model or instruction-data scale, but also on how geospatial evidence is represented, supervised, verified, and aligned with task-specific constraints.

7.1. Data

Data remains the primary bottleneck for remote sensing reasoning. Unlike conventional perception tasks, reasoning requires complex supervision that goes beyond final answers, including question construction, spatial grounding, logical evidence, and intermediate signals such as masks or temporal relations. Recent works have introduced increasingly rich resources, such as segmentation-aware VQA in *EarthVL* [113], pixel-level reasoning in *SegEarth-R1* [114], language-guided segmentation in *SegEarth-R2* [115], and multi-modal reasoning traces in *TerraScope* [26], as well as reasoning-oriented corpora in *GeoReason* [53], *RSThinker* [121], and *GeoSolver* [54]. The 2026 Q2 additions make scale interpretation more important: Delta-QA derives many prompts from annotated change observations, APE-GRO filters synthetic tool-interleaved trajectories, GeoX generates executable programs over unlabeled

 doi:10.20944/preprints202608.1840.v1

image pools, and MMRS-OneVision converts or synthesizes a very large instruction collection from public resources [40,44,46,47].

This reliance leads to several structural limitations. First, synthetic reasoning often fails to reflect expert-level geospatial analysis, especially in long-horizon or uncertain scenarios. Second, annotations are typically constructed per image or task, lacking consistent alignment across multi-temporal and multi-source data. Third, high-quality annotation is inherently expensive because it requires domain expertise in areas such as urban analysis, ecology, and disaster science. As discussed in *GeoZero* [118], even carefully designed reasoning traces may introduce bias and constrain reasoning diversity.

We therefore report scale by supervision unit rather than computing a single aggregate. Four unit families recur in the literature: (1) underlying observations and dense annotations, such as images and masks; (2) language-level supervision, such as question–answer, instruction, or text pairs; (3) process or optimization supervision, such as trajectories, reasoning steps, and reward samples; and (4) agentic evaluation units, such as tasks or episodes. Counts from different families are not additive: a single image may generate many correlated prompts, one task may contain many tool steps, and a process-reward sample is not equivalent to an independently acquired observation. Even counts within the same family should not be pooled when dataset overlap, generation pipelines, geographic coverage, or train/test reuse are undisclosed. Table 7 consequently preserves the terminology and unit reported by each paper, with no cross-paper total or percentage.

Table 7. Reasoning-related data scales reported by representative papers, separated by the unit named in each source. SFT denotes supervised fine-tuning and RL denotes reinforcement learning. K/M values are approximate. No cross-unit or cross-paper aggregate is computed.

Work	Dataset / corpus	Reported scale, separated by unit	Counted unit family / role
Supervised reasoning			
EarthVL [113]	EarthVLSet	Images: 10.9K; text pairs: 761.6K	Observations; language supervision
SegEarth-R1 [114]	EarthReason	Image–mask pairs: 5.4K; implicit QA pairs: >30K	Dense annotations; language supervision
SegEarth-R2 [115]	LaSeRS	Masks: 40K; QA pairs: 31K	Dense annotations; language supervision
TerraScope [26]	Terra-CoT / TerraScope-Bench	Training samples: 1.0M; benchmark samples: 3.8K	Paper-defined training and evaluation instances
Delta-LLaVA [40]	Delta-QA	QA pairs: 180,876; bi- and tri-temporal source tasks	Language and dense temporal supervision
RL-driven reasoning			
RS-EoT [122]	RS-EoT-4K	SFT traces: 4.3K; two subsequent RL stages: source datasets reported, comparable instance count not reported	Language/process supervision
Geo-R [128]	MP16-Rand-500K / MP16-Hard-200K	Supervised samples: 500K; reward-based samples: 200K	Language supervision; optimization supervision
Geo-R1 [119]	Few-shot splits	Task-specific examples: 26/130/260; 20/40; 24/120/240	Few-shot training examples across separate tasks
GeoReason [53]	GeoReason-Bench	Trajectories: 4K (1K supervised; 3K reward-based)	Process trajectories; optimization supervision
GeoVLM-R1 [117]	Aggregated EO QA pairs	QA pairs: 702K	Language supervision
GeoZero [118]	GeoZero-Raw / Instruct / Hard	Raw examples: 755K; supervised instructions: 610K; reward-based samples: 20K	Raw pool; language supervision; optimization supervision
RSThinker [121]	Geo-CoT380k	Structured SFT rationales: 384,591; additional RL inputs: 110,790; GRPO also reuses rationale-free Geo-CoT inputs	Language/process supervision; optimization supervision
GeoSolver [54]	Geo-PRM-2M	Process-supervision samples: 2.0M	Process / reward-model supervision
RemoteAgent [39]	VagueEO	Train pairs: 5,000 (five intrinsic tasks); test pairs: 1,000 (ten tasks, including five task types unseen during RL)	Sparse-task RL pairs; task-routing evaluation
RemoteReasoner [123]	EarthReason-derived prompts	Training images: 2.4K; prompts: 14.2K (six questions per image)	Observations; correlated language supervision
TinyRS-R1 [120]	VHM-based pipeline	Pretraining samples: 1M; instruction samples: 100K; reward-based samples: 10K	Pretraining; language supervision; optimization supervision
GeoVista [44]	APE-GRO	Raw trajectories: 142,125; validated trajectories: 33,445; RL restricted to grounding and counting	Process trajectories; optimization subset
GeoX [46]	SAMRS-FAST / Globe230k	Source images: 64,147 / 232,819; self-play programs: ∼6.5K	Unlabeled image pools; executable programs
Agentic / tool-augmented reasoning			
Earth-Agent [28]	Earth-Bench	Tasks: 248; images: 13.7K	Agentic tasks; supporting observations
EarthAgent [124]	GeoPlan-bench	Planning tasks: 1,244	Agentic planning tasks
OpenEarthAgent [37]	OpenEarthAgent dataset	Train/evaluation instances: 14,538/1,169; train/evaluation reasoning steps: 100,656/7,064	Task instances; process steps, separated by split
ThinkGeo [127]	ThinkGeo benchmark	Tasks: 486; expert-verified steps: 1,778; executable tools: 14	Agentic tasks; process steps
GeoMMAgent [38]	GeoMMBench	Expert multiple-choice questions: 1,053	Evaluation questions
MAP-Agent [43]	UHR-Micro	Images: 1,212; instructions: 11,253; tasks: 16	Observations; agentic evaluation instructions
TerraAgent [48]	TerraBench	Tasks: 403; verified steps: ∼24.5K; sub-tools: 77	Agentic tasks; process steps; tool environment

Accordingly, the strongest supported conclusion is a diversity, provenance, and validity gap rather than a universal shortage of nominal samples. Future datasets should disclose unique source observations, geographic and temporal coverage, sensor composition, generation and filtering procedures, annotator expertise, and correlations between derived instructions. Authentic multi-temporal and uncertainty-aware supervision remains especially important for measuring generalization beyond synthetic reasoning traces.

This gap should be tested with source-observation-disjoint, geography-disjoint, and time-disjoint splits rather than random prompt splits alone. A useful data audit should report the number of

 doi:10.20944/preprints202608.1840.v1

prompts derived from each observation, compare synthetic and expert-authored traces under blinded review, and measure performance separately across geography, sensor, season, and task composition. Progress is then an improvement under controlled evidence coverage, not merely an increase in the number of correlated instructions.

7.2. Model

To avoid a second, competing taxonomy, we retain the two axes in Section 5 and treat evidence representation, policy optimization, output granularity, and execution as cross-cutting diagnostic dimensions. For each dimension, the important question is not which method family a system resembles, but which observable failure it exhibits, which intervention exposes that failure, and what result would constitute progress.

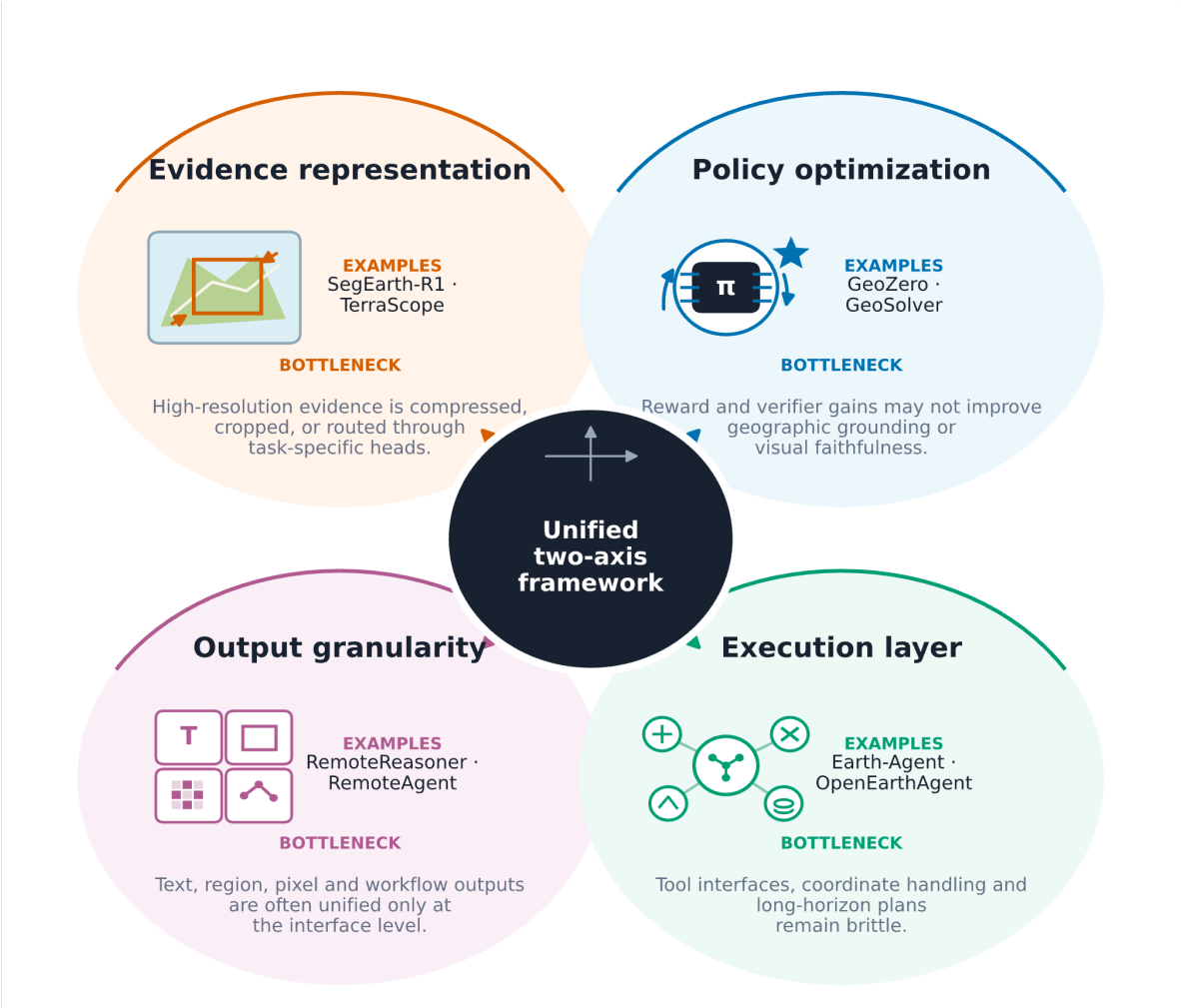


Figure 8. Cross-cutting model-design bottlenecks under the unified two-axis framework. The four illustrated dimensions characterize evidence representation, policy optimization, output granularity, and the execution layer. Representative systems are shown for each dimension, together with their primary bottlenecks. These dimensions provide complementary diagnostic perspectives rather than defining an alternative four-family taxonomy.

Evidence Representation: Preserving Decisive Observations

SegEarth-R1 and *TerraScope* show the value of making regions or masks observable, while *RS-EoT* identifies a narrower failure mode in which a model constructs a long rationale after one coarse visual pass [26,114,122]. These results do not establish that a compressed visual token sequence preserves every small object, boundary, or long-range relation needed by a later conclusion. The appropriate diagnostic is a controlled degradation study: vary input resolution, remove or occlude the cited region, drop one modality, or shuffle temporal order, and measure both task performance

 doi:10.20944/preprints202608.1840.v1

and evidence localization. A model demonstrates stronger evidence preservation only if its decisions change predictably with decisive evidence and remain stable under irrelevant perturbations.

UHR-Micro isolates this problem by making the decisive target occupy less than one percent—and often less than 0.01%—of the image, while MAP-Agent selectively revisits candidate regions at native resolution [43]. SkyNative's RSME-Bench complements this with progressive visual degradation and misleading-text tests [45]. These experiments diagnose micro-evidence access and visual reliance, respectively; they do not by themselves establish multi-step faithfulness. GeoVista exposes a richer but still bounded trace through plans, crop calls, and an evidence state [44]. A necessary next test is to perturb the selected region while preserving the question and global scene, then measure whether the answer and evidence state change for the right reason.

Policy Optimization: Separating Better Outputs from Better Evidence Use

Reward-based post-training can improve a selected behavior without showing that the policy encodes stronger geographic structure. For example, RSThinker applies outcome-based GRPO after rationale supervision, GeoReason tests conclusion stability under option permutation, and GeoSolver learns token-level process scores [53,54,121]. These signals operate at different credit granularities and should not be treated as equivalent proof of faithful reasoning. A convincing optimization study should therefore include evidence ablations, verifier–expert agreement, and reward-component ablations, and should test whether a policy still succeeds when linguistic shortcuts are preserved but the relevant image, location, or time evidence is altered.

RemoteZero, GeoVista, GeoX, and RS-HyRe-R1 sharpen this distinction through crop-semantic, state-machine plan, executable-program, and hybrid visual–semantic rewards, respectively [41,42, 44,46]. Each verifier observes a different slice of behavior and can be exploited differently: a crop score can favor oversized regions, a plan score can reward formal state transitions without evidential sufficiency, and program execution can inherit segmentation errors. Claims of improved reasoning should therefore report verifier failures as well as policy gains.

Output Granularity: Interface Unification Versus Shared Representation

RemoteReasoner predicts regions with the MLLM and derives masks and contours through segmentation and morphological transformations, whereas *RemoteAgent* resolves image- and sparse-region tasks internally but routes dense requests to specialist MCP tools [39,123]. They therefore unify task access and output orchestration rather than learning one shared representation that directly produces every granularity. This distinction can be tested by holding observations and training data fixed, switching the required output unit, and checking cross-output consistency: a textual relation, box, mask, and derived measurement should agree geometrically and semantically without task-specific repair.

Delta-LLaVA directly couples change text and dense masks, whereas Earth-OneVision tokenizes text, coordinates, and downsampled run-length-encoded masks within a unified autoregressive interface [40,47]. The former is specialized to short temporal sequences; the latter broadens task and sensor coverage but its compact mask encoding imposes a spatial-detail ceiling. Output unification should therefore be tested through cross-output agreement and boundary-sensitive metrics, not inferred from a common decoder alone.

Execution Layer: Tool Use Under Recoverable Failure

Earth-Agent exposes an inference-time ReAct controller over EO tools, while *OpenEarthAgent* learns next-tool behavior from replay-validated trajectories [28,37]. Their success makes planning and tool state observable, but also creates new failure surfaces in schemas, coordinates, arguments, returned artifacts, and external service state. Execution reliability should be measured by injecting coordinate offsets, empty results, stale observations, timeouts, and contradictory tool outputs, then reporting detection, recovery, and final-task degradation. Completing an unperturbed workflow is not sufficient evidence of robust long-horizon execution.

 doi:10.20944/preprints202608.1840.v1

TerraAgent adds gridded environmental retrieval, forecasting, simulation, and provenance-bearing artifacts to this execution surface, while GeoX uses a much narrower deterministic geometry library plus one open-vocabulary segmenter [46,48]. These systems should not be compared by tool count alone. TerraAgent requires fault recovery and numerical validation across heterogeneous services; GeoX requires sensitivity analysis to the segmenter's masks and the expressiveness of its primitive set.

A shared Diagnostic Protocol

The main model-level failures can be reduced to three testable chains: evidence loss can trigger a hallucination cascade, an invalid tool result can propagate through later actions, and distribution shift can produce confident errors. These chains should be tested with resolution and evidence interventions, tool-fault injection, and geography-, sensor-, season-, and time-shifted evaluation. Progress should be reported through degradation curves, error-recovery rate, calibration or selective risk, and comparison with a perception-plus-GIS baseline under matched data access.

7.3. Evaluation and Scientific Validity

A reasoning claim should be matched to the evidence layer that can actually support it. Outcome scores establish task performance, spatial metrics establish output localization, exposed trajectories establish properties of visible actions, and repeated sampling establishes stability. None of these measurements alone proves that the complete reasoning chain is faithful. Evaluation should therefore begin with the capability being claimed and specify the intervention that could falsify that claim.

Outcome, Grounding, and Diagnostic Proxies

Conventional task metrics used by *EarthVL*, *GeoVLM-R1*, and *TinyRS-R1* measure whether the requested output is correct [113,117,120]. Region and mask metrics in *SegEarth-R1/2*, *RemoteReasoner*, and *TerraScope* add spatial evidence, but a correct mask still does not show that the model used it causally [26,114,115,123]. Similarly, GeoReason's option-permutation test is a conclusion-stability proxy, while the Avg@5, Conv@5, and Pass@5 values reported by RS-EoT characterize correctness and stability across five samples; they are evaluation metrics, not training rewards or verifiers of individual steps [53,122]. Such diagnostics are useful when their scope is stated explicitly.

Plans, Trajectories, and Live Execution

Agentic evaluation should distinguish a proposed plan from an executed workflow. *EarthAgent* reports key-tool recall, precision, and F1, normalized path similarity, and a judge-based score for plan quality independently of actual tool execution [124]. *Earth-Agent* evaluates reference trajectories and final outcomes, while *OpenEarthAgent* separates step-wise action, tool, and argument quality from live end-to-end tool-set or order fidelity and final-task accuracy [28,37]. *ThinkGeo* is an evaluation benchmark rather than a trained agent; it measures instruction, tool, argument, and summary accuracy step by step, together with tool-category and final-answer accuracy end to end [127]. These protocols expose different failure points and should not be collapsed into one generic trajectory score.

The Q2 systems add four distinct evaluation granularities. GeoMMAgent is assessed primarily through final multiple-choice accuracy; MAP-Agent reports task scores and the cost of repeated visual calls; RemoteAgent reports intent and downstream task performance without a learned tool-trajectory metric; and TerraBench separates instruction, tool, argument, order, and execution measures from tolerance-aware numerical outcomes [38,39,43,48]. Their results are therefore complementary, not directly rankable. In particular, TerraBench demonstrates that a high process score need not imply a correct final number.

Transfer, Calibration, and Resource Use

Robustness-oriented evidence is heterogeneous rather than interchangeable. *Geo-R* supplies adjacent cross-benchmark geolocation evidence on ground-level imagery, *Geo-R1* reports cross-dataset

 doi:10.20944/preprints202608.1840.v1

few-shot transfer, and *RemoteReasoner* tests held-out tasks and a small set of categories absent from its training corpus [119,123,128]. These results probe dataset, task, or category transfer, but do not constitute controlled robustness across remote-sensing geography, sensor, season, resolution, and acquisition time. *VRA* should instead be read as an inference-time accuracy–cost study on a 500-question VRSBench subset; it does not evaluate those distribution shifts [126].

Minimum Reporting Protocol

A credible RS-Reasoning benchmark should jointly report: (1) task correctness; (2) spatial or multimodal grounding; (3) process or tool-trajectory validity; (4) calibration and selective prediction, including performance after abstaining on uncertain cases; (5) robustness across geography, sensor, season, resolution, and acquisition time; and (6) computational and human-review cost. Splits should be constructed by geographic region and acquisition time, rather than by random image alone, to reduce near-duplicate and location leakage. Results should include confidence intervals or repeated-run variation, and verifier-based scores should be audited against expert judgments on a documented subset. Table 8 converts these reporting layers into falsifiable evaluation requirements.

Table 8. Capability-to-evidence matrix for remote sensing reasoning. Each claim is paired with a required intervention and a shortcut that would make the evidence insufficient.

Claimed capability	Required evidence	Required split or intervention	Invalid shortcut
Task competence	Outcome metric matched to the task, with uncertainty or repeated-run variation	Geography- and time-disjoint test data; source-observation deduplication	Random prompt split with near-duplicate imagery
Grounded reasoning	Correct answer plus localized visual, spectral, temporal, or GIS evidence	Remove, occlude, or counterfactually replace the claimed evidence	Plausible rationale or mask overlap without evidence-use testing
Process or execution validity	Verified consequential steps, tool arguments, returned values, and final output	Inject an invalid step, coordinate, empty result, timeout, or contradictory observation	Correct final answer obtained by luck or an invalid trajectory
Robustness and calibration	Degradation and selective-risk curves under documented shifts	Hold out geography, sensor, season, resolution, category, and acquisition period	Average in-domain score or a single small unseen-category test
Operational utility	Latency, memory, token and tool cost, abstention, and human-review time	Compare with a strong perception-plus-GIS pipeline under matched data access	Unbounded search or tool budget without a matched baseline

Table 9. Verifier targets for auditable remote sensing reasoning. These are proposed evaluation and optimization requirements, not capabilities already implemented by every cited system.

Verifier target	Auditable signal	Exploit or failure test
Evidence and geographic constraints	Region overlap; containment, adjacency, distance, direction, topology; cross-modal or temporal agreement	Remove decisive evidence, swap time order, or introduce a plausible but invalid geographic relation
Process and tool execution	Checkable intermediate claims; tool schema, argument, returned value, and state-transition validity	Inject a wrong coordinate, unsupported step, stale return, or silent tool failure
Uncertainty and resource use	Calibrated correctness, abstention quality, latency, tool calls, and human-review cost	Selective-risk and fixed-budget evaluation against a matched baseline

7.4. Training Objectives and Verifiers

Current optimization signals differ in what they can observe. RSThinker derives its GRPO reward from canonical final-output metrics after structured rationale SFT, while GeoVLM-R1 uses output-specific task rewards; RS-EoT applies separate grounding and VQA rewards, and RemoteAgent dispatches reward functions by sparse output type [39,117,121,122]. GeoZero adds an answer-anchored thinking proxy, GeoReason tests conclusion stability under option permutation, and GeoSolver moves credit inside the trajectory through a learned token-level process scorer [53,54,118]. RemoteZero scores

 doi:10.20944/preprints202608.1840.v1

crop–query compatibility, GeoVista adds a state-machine plan proxy to endpoint rewards, GeoX uses program execution for self-play, and RS-HyRe-R1 combines several visual–semantic signals [41,42,44,46]. GeoMMAgent, MAP-Agent, and TerraAgent are inference-time orchestration frameworks rather than examples of policy optimization [38,43,48]. These objectives span endpoint scoring, trajectory-level proxies, and learned within-trajectory feedback; answer accuracy, IoU, semantic similarity, or output format can verify an observable result without establishing that the intermediate rationale caused it, so they should not all be described as process verification.

A useful design target is to report a vector of verifier outputs before reducing them to one scalar:

$$\begin{aligned} \mathbf{r} &= (R_{\text{ans}}, R_{\text{ground}}, R_{\text{constraint}}, \\ &\quad R_{\text{process}}, R_{\text{calibration}}, -C_{\text{tool}}), \\ R_{\lambda} &= \boldsymbol{\lambda}^{\mathsf{T}}\mathbf{r}. \end{aligned} \tag{4}$$

The components represent final correctness, observable evidence, geographic constraints, checkable steps, calibrated confidence, and execution cost or failure risk. Scalar aggregation is appropriate only when the weights and normalization are reported and ablated; otherwise a high answer score can silently compensate for invalid grounding or execution. Each component also requires a held-out audit because combining imperfect verifiers can create a more complex but still exploitable objective.

From Proxy Scores to Independently Checkable Claims

Process reward models (PRMs) assign credit to intermediate steps rather than only final outcomes [133]; GeoSolver applies this idea through learned token-level geospatial process feedback [54]. Such learned feedback can inherit annotation or model bias and is therefore not a ground-truth checker. RSThinker illustrates the remaining boundary: it supervises structured rationales during SFT, yet its RL reward is determined by final task metrics. GeoReason's option-permutation stability is likewise a trajectory-level diagnostic rather than verification of every geographic statement [53,121]. Remote sensing therefore needs verifiers matched to claim type: geometric operators for regions, topology for spatial relations, ordering and persistence tests for temporal claims, cross-sensor checks for multimodal evidence, and schema plus returned-value validation for tools.

Minimum Verifier Audit

Every reward or verifier component should undergo three tests: agreement with blinded expert judgment on a documented subset, component ablation to identify which behavior it changes, and adversarial or distribution-shift testing to expose reward hacking. Expert comparisons can also supply preference data for evidential sufficiency or operational usefulness, but they should complement rather than replace reproducible geographic checks. The success criterion is not a longer chain of thought, but a trajectory whose consequential claims are independently checkable against geographic evidence, with fewer unsupported claims, better calibrated abstention, and improved task performance under a fixed evidence and execution budget.

7.5. Reproducibility, Efficiency, and Governance

Reproducibility and System Dependence

Reasoning results can depend on model and data versions, prompt templates, sampling parameters, verifier implementations, tool schemas, external service state, and the order in which geospatial operations are executed. Reproducible reporting should therefore include versioned prompts and tool registries, random seeds, geographic split construction, failure and retry policies, and enough intermediate artifacts to replay representative trajectories. For agentic systems, final-answer accuracy without executable traces or tool-state records is insufficient to isolate whether an error arose from perception, planning, arguments, software, or external data.

 doi:10.20944/preprints202608.1840.v1

Efficiency and Baseline Parity

Reasoning systems should be compared with strong perception-plus-GIS pipelines under matched data access and output requirements. Reports should separate visual-token cost, generated-token cost, tool calls, latency, memory, and human correction time, because a higher task score may be offset by expensive test-time search or repeated tool execution. MAP-Agent, for example, averages about 3.53 VLM calls per question under its selected protocol, while GeoX uses 32 sampled solutions and majority voting for evaluation [43,46]. Such budgets must accompany accuracy and cannot be treated as architecture-independent efficiency. Efficiency is therefore part of validity for operational EO workflows, not merely an implementation detail.

Governance, Privacy, and Dual Use

Operational deployment also requires explicit treatment of data licensing and provenance, sensitive-location exposure, privacy risks from linking imagery with auxiliary records, dual-use analysis, access control for external tools, and the allocation of final decision authority. A comprehensive synthesis of these governance topics requires additional domain-specific references and should not be inferred from the technical benchmark papers currently reviewed. Until that literature is incorporated, the survey should present governance as an identified coverage gap rather than a resolved design framework.

8. Conclusion and Future Directions

8.1. Conclusion

The evidence reviewed in this survey supports a narrower conclusion than the rhetoric of autonomous geospatial reasoning sometimes suggests. Remote sensing systems are beginning to combine visual observations, spatial outputs, optimization signals, and external tools, but the resulting capabilities remain uneven and are usually demonstrated on task-specific benchmarks. The three paradigms reviewed here are therefore best understood as complementary mechanisms for acquiring or executing reasoning: supervision supplies explicit evidence structures, reinforcement learning shapes behavior under verifiers, and agents expose executable decisions. None of these mechanisms, by itself, guarantees that a conclusion is grounded, calibrated, or operationally reliable.

The field's central methodological shift should be from evaluating persuasive outputs to testing falsifiable evidence use. A reasoning claim is strongest when decisive observations can be localized, intermediate constraints can be checked, tool actions can be replayed, uncertainty supports abstention, and performance survives geography-, sensor-, and time-disjoint evaluation under a reported resource budget. This standard connects the survey's definition, taxonomy, and application evidence: progress is not a longer rationale or a more elaborate controller, but a geospatial conclusion that an analyst can inspect, challenge, and recompute.

8.2. Future Directions

Future work should be organized around research questions whose answers can be rejected empirically. Rather than assigning speculative calendar horizons, Table 10 maps five coupled questions to an intervention path and a predefined success standard. These directions are not additional model families; they identify the missing evidence needed to move from benchmark demonstrations toward dependable geospatial analysis.

The roadmap exposes dependencies rather than a linear sequence: representation and retrieval determine evidence quality, verifiers shape learned behavior, execution turns claims into actions, and deployment studies test whether benefits survive human and institutional constraints. A strong future study should therefore name the failure it targets, implement an intervention specific to that failure, and state both a success measure and a result that would falsify the claim. This practice would turn future directions from a list of promising components into a cumulative research program.

 doi:10.20944/preprints202608.1840.v1

Table 10. A falsifiable roadmap for remote sensing reasoning. Each research question is paired with an intervention path and evidence that would constitute progress.

Research question	Intervention path	Success standard
Can decisive local evidence and global spatial structure be preserved at operational image scales?	Budget-matched multi-scale encoding, region memory, cross-tile spatial coordinates, and evidence occlusion or replacement tests. Selective-token and active-zoom systems provide current starting points [134–137].	Joint task and grounding gains on geography-disjoint data; predictable degradation when decisive evidence is removed; stability under irrelevant perturbations with matched compute.
Can external geospatial knowledge improve inference without adding stale or untraceable facts?	Provenance- and time-aware retrieval over raster, vector, topology, and domain records; explicit source links; retrieval ablations and outdated-record tests. OSM-derived supervision and location embeddings are precedents, while inference-time graph retrieval remains a proposal [84,87].	Fewer geographic constraint violations and unsupported claims; gains disappear when relevant records are withheld; stale or conflicting records trigger detection, qualification, or abstention.
Can reasoning models execute and repair GIS workflows rather than only propose plausible plans?	Typed tool schemas, precondition and postcondition checks, persistent execution state, replay logs, and fault injection. Current EO agents establish tool-orchestration starting points but not robust recovery [27,28,37].	Higher executable-trajectory and final-task accuracy; better recovery under injected coordinate, schema, timeout, empty-return, or contradictory-result faults; lower cost and expert correction time against a matched workflow baseline.
Can scalable supervision teach geospatial reasoning without scaling synthetic shortcuts?	Prospective self- or weakly supervised spatial, temporal, and cross-modal proxy tasks derived from observable geospatial invariants; programmatic verifiers; expert-audited subsets; generator and verifier ablations.	High verifier–expert agreement, transfer to expert-authored and geography-disjoint tests, and fewer unsupported consequential claims rather than gains restricted to synthetic data.
When does an RS-Reasoning system improve real decisions without increasing high-consequence risk?	Prospective analyst studies with calibrated abstention and escalation, provenance and access controls, scenario-based safety tests, and fixed evidence and resource budgets. Published systems remain benchmark or scenario-based demonstrations rather than prospective deployment evidence [27,28,37].	Reduced review or correction time with non-inferior or better decision quality; calibrated selective risk; documented failure recovery; no increase in prespecified severe errors.

References

1. Q. Ma, J. Pan, and C. Bai, "Direction-oriented visual–semantic embedding model for remote sensing image–text retrieval," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 62, pp. 1–14, 2024.
2. T. C. Vance, T. Huang, and K. A. Butler, "Big data in earth science: Emerging practice and promise," *Science*, vol. 383, no. 6688, p. eadh9607, 2024.
3. L. Filchev, L. Pashova, V. Kolev, and S. Frye, "Challenges and solutions for utilizing earth observations in the" big data" era," *arXiv preprint arXiv:2108.08886*, 2021.
4. Q. Zhao, L. Yu, Z. Du, D. Peng, P. Hao, Y. Zhang, and P. Gong, "An overview of the applications of earth observation satellite data: impacts and future trends," *Remote Sensing*, vol. 14, no. 8, p. 1863, 2022.
5. J. Pan, Q. Ma, and C. Bai, "Reducing semantic confusion: Scene-aware aggregation network for remote sensing cross-modal retrieval," in *Proceedings of the 2023 ACM International Conference on Multimedia Retrieval*, 2023, pp. 398–406.
6. X. He, C. Tang, X. Liu, W. Zhang, Z. Gao, C. Li, S. Qiu, and J. Xu, "Spectral discrepancy and cross-modal semantic consistency learning for object detection in hyperspectral images," *IEEE Transactions on Multimedia*, vol. 27, pp. 6719–6731, 2025. [Online]. Available: https://doi.org/10.1109/TMM.2025.3586155
7. D. Yu and C. Fang, "Urban remote sensing with spatial big data: A review and renewed perspective of urban studies in recent decades," *Remote Sensing*, vol. 15, no. 5, p. 1307, 2023.
8. S. A. H. Mohsan, M. A. Khan, F. Noor, I. Ullah, and M. H. Alsharif, "Towards the unmanned aerial vehicles (uavs): A comprehensive review," *Drones*, vol. 6, no. 6, p. 147, 2022.
9. Z. Zou, G. Sun, Z. Wei, J. Pan, Y. Li, M. Peng, and W. Xu, "Self in space: Benchmarking self-awareness and spatial cognition in uav embodied intelligence," *arXiv preprint arXiv:2607.12477*, 2026.
10. A. Alexopoulos, K. Koutras, S. B. Ali, S. Puccio, A. Carella, R. Ottaviano, and A. Kalogeras, "Complementary use of ground-based proximal sensing and airborne/spaceborne remote sensing techniques in precision agriculture: A systematic review," *Agronomy*, vol. 13, no. 7, p. 1942, 2023.
11. X. Li, C. Wen, Y. Hu, Z. Yuan, and X. X. Zhu, "Vision-language models in remote sensing: Current progress and future trends," *IEEE Geoscience and Remote Sensing Magazine*, vol. 12, no. 2, pp. 32–66, 2024.
12. A. Xiao, W. Xuan, J. Wang, J. Huang, D. Tao, S. Lu, and N. Yokoya, "Foundation models for remote sensing and earth observation: A survey," *IEEE Geoscience and Remote Sensing Magazine*, 2025.
13. S. Zhao, K. Tu, S. Ye, H. Tang, Y. Hu, and C. Xie, "Land use and land cover classification meets deep learning: A review," *Sensors*, vol. 23, no. 21, p. 8966, 2023.

 doi:10.20944/preprints202608.1840.v1

14. M. A. Moharram and D. M. Sundaram, "Land use and land cover classification with hyperspectral data: A comprehensive review of methods, challenges and future directions," *Neurocomputing*, vol. 536, pp. 90–113, 2023.
15. Z. Zhao, F. Islam, L. A. Waseem, A. Tariq, M. Nawaz, I. U. Islam, T. Bibi, N. U. Rehman, W. Ahmad, R. W. Aslam *et al.*, "Comparison of three machine learning algorithms using google earth engine for land use land cover classification," *Rangeland ecology & management*, vol. 92, pp. 129–137, 2024.
16. X. He, C. Tang, X. Liu, W. Zhang, K. Sun, and J. Xu, "Object detection in hyperspectral image via unified spectral-spatial feature aggregation," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 61, pp. 1–13, 2023. [Online]. Available: https://doi.org/10.1109/TGRS.2023.3307288
17. X. He, C. Tang, X. Zou, and W. Zhang, "Multispectral object detection via cross-modal conflict-aware learning," in *Proceedings of the 31st ACM International Conference on Multimedia*. ACM, Oct. 2023, pp. 1465–1474. [Online]. Available: https://doi.org/10.1145/3581783.3612651
18. S. Gui, S. Song, R. Qin, and Y. Tang, "Remote sensing object detection in the deep learning era—a review," *Remote Sensing*, vol. 16, no. 2, p. 327, 2024.
19. J. Pan, Y. Liu, X. He, L. Peng, J. Li, Y. Sun, and X. Huang, "Enhance then search: An augmentation-search strategy with foundation models for cross-domain few-shot object detection," in *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025, pp. 1548–1556.
20. J. Pan, Y. Liu, Y. Fu, M. Ma, J. Li, D. P. Paudel, L. Van Gool, and X. Huang, "Locate anything on earth: Advancing open-vocabulary object detection for remote sensing community," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 39, no. 6, 2025, pp. 6281–6289.
21. J. Li, J. Pan, Y. Sun, and X. Huang, "Semantic-aware ship detection with vision-language integration," in *IGARSS 2025-2025 IEEE International Geoscience and Remote Sensing Symposium*. IEEE, 2025, pp. 6903–6907.
22. L. Huang, B. Jiang, S. Lv, Y. Liu, and Y. Fu, "Deep-learning-based semantic segmentation of remote sensing images: A survey," *IEEE journal of selected topics in applied earth observations and remote sensing*, vol. 17, pp. 8370–8396, 2023.
23. X. Ma, X. Zhang, M.-O. Pun, and M. Liu, "A multilevel multimodal fusion transformer for remote sensing semantic segmentation," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 62, pp. 1–15, 2024.
24. Y. Liu, Y. Zhang, Y. Wang, and S. Mei, "Rethinking transformers for semantic segmentation of remote sensing images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 61, pp. 1–15, 2023.
25. S. Akcali and A. Ispalar Cahantimur, "How socio-spatial aspects of urban space influence social sustainability: a case study," *Journal of housing and the built environment*, vol. 38, no. 4, pp. 2525–2557, 2023.
26. Y. Shu, B. Ren, Z. Xiong, X. X. Zhu, B. Demir, N. Sebe, and P. Rota, "TerraScope: Pixel-grounded visual reasoning for earth observation," *arXiv preprint arXiv:2603.19039*, 2026.
27. A. Bell, A. Aides, A. Helmy, A. Muslim, A. Barzilai, A. Slobodkin, B. Jaber, D. Schottlander, G. Leifman, J. Paul *et al.*, "Earth AI: Unlocking geospatial insights with foundation models and cross-modal reasoning," *arXiv preprint arXiv:2510.18318*, 2025.
28. P. Feng, Z. Lv, J. Ye, X. Wang, X. Huo, J. Yu, W. Xu, W. Zhang, L. Bai, C. He, and W. Li, "Earth-Agent: Unlocking the full landscape of earth observation with agents," in *International Conference on Learning Representations*, 2026. [Online]. Available: https://openreview.net/forum?id=dkIXAbWuxO
29. L. Bashmal, Y. Bazi, F. Melgani, M. M. Al Rahhal, and M. A. Al Zuair, "Language integration in remote sensing: Tasks, datasets, and future directions," *IEEE Geoscience and Remote Sensing Magazine*, vol. 11, no. 4, pp. 63–93, 2023. [Online]. Available: https://doi.org/10.1109/MGRS.2023.3316438
30. X. Weng, C. Pang, and G.-S. Xia, "Vision-language modeling meets remote sensing: Models, datasets and perspectives," *IEEE Geoscience and Remote Sensing Magazine*, 2025, early Access.
31. C. Liu, J. Zhang, K. Chen, M. Wang, Z. Zou, and Z. Shi, "Remote sensing spatiotemporal vision–language models: A comprehensive survey," *IEEE Geoscience and Remote Sensing Magazine*, vol. 14, no. 1, pp. 383–423, 2026. [Online]. Available: https://doi.org/10.1109/MGRS.2025.3598283
32. Y. Wang, W. Chen, X. Han, X. Lin, H. Zhao, Y. Liu, B. Zhai, J. Yuan, Q. You, and H. Yang, "Exploring the reasoning abilities of multimodal large language models (mllms): A comprehensive survey on emerging trends in multimodal reasoning," *arXiv preprint arXiv:2401.06805*, 2024.
33. Q. Ma, C. Qiu, X. Cheng, X. Zhang, P. Duan, K. Yang, X. Kang, and S. Li, "Multimodal large language models for remote sensing image understanding: Domain-specific or general-purpose?" *arXiv preprint arXiv:2607.20284*, 2026.

 doi:10.20944/preprints202608.1840.v1

34. W. Zhang, M. Cai, T. Zhang, Y. Zhuang, and X. Mao, "Earthgpt: A universal multimodal large language model for multisensor image comprehension in remote sensing domain," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 62, pp. 1–20, 2024.
35. K. Kuckreja, M. S. Danish, M. Naseer, A. Das, S. Khan, and F. S. Khan, "Geochat: Grounded large vision-language model for remote sensing," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 27 831–27 840.
36. J. Luo, Z. Pang, Y. Zhang, T. Wang, L. Wang, B. Dang, J. Lao, J. Wang, J. Chen, Y. Tan, and Y. Li, "Skysensegpt: A fine-grained instruction tuning dataset and model for remote sensing vision-language understanding," *arXiv preprint arXiv:2406.10100*, 2024.
37. A. Shabbir, M. U. Sheikh, M. A. Munir, H. Debary, M. Fiaz, M. Z. Zaheer, P. Fraccaro, F. S. Khan, M. H. Khan, X. X. Zhu *et al.*, "OpenEarthAgent: A unified framework for tool-augmented geospatial agents," in *European Conference on Computer Vision*, 2026. [Online]. Available: https://arxiv.org/abs/2602.17665
38. A. Xiao, S. Cheng, Y. Xu, Y. Ren, H. Chen, and N. Yokoya, "GeoMMBench and GeoMMAgent: Toward expert-level multimodal intelligence in geoscience and remote sensing," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2026, pp. 34 843–34 853.
39. L. Yao, S. Xu, F. Liu, C. Zhang, B. Yao, R. Min, Y. Li, C. Ouyang, S. Di, and M.-L. Zhang, "RemoteAgent: Bridging vague human intents and earth observation with rl-based agentic mllms," 2026. [Online]. Available: https://arxiv.org/abs/2604.07765
40. X. Li, J. Li, K. Zhang, Y. Fang, L. Lin, H. Wang, H. Wu, and Z. Fan, "Decoding the delta: Unifying remote sensing change detection and understanding with multimodal large language models," 2026. [Online]. Available: https://arxiv.org/abs/2604.14044
41. G. Zhou, H. He, P. Shen, J. Zhang, L. Zhang, L. Xu, Z. Wang, Z. Li, X. Cui, W. Guo, and H. Li, "RS-HyRe-R1: A hybrid reward mechanism to overcome perceptual inertia for remote sensing images understanding," 2026. [Online]. Available: https://arxiv.org/abs/2604.17504
42. L. Yao, F. Liu, S. Xu, C. Zhang, R. Min, S. Di, and Y. Zheng, "RemoteZero: Geospatial reasoning with zero labels," 2026. [Online]. Available: https://arxiv.org/abs/2605.04451
43. S. Ni, T. Wang, J. Zhang, H. Chen, H. Guo, N. Zhang, and B. Du, "UHR-Micro: Diagnosing and mitigating the resolution illusion in earth observation VLMs," 2026. [Online]. Available: https://arxiv.org/abs/2605.12237
44. J. Zhu, R. Fu, J. Hu, J. Huang, N. Xing, and B. Yang, "GeoVista: Visually grounded active perception for vision-language understanding of ultra-high-resolution remote sensing images," 2026. [Online]. Available: https://arxiv.org/abs/2605.14475
45. X. Yang, R. Fu, Z. Lin, Z. Duan, J. Zhu, J. Hu, L. Sun, W. Zhang, J. Liu, X. Na, H. Liu, W. Zhang, and B. Yang, "SkyNative: A native multimodal framework for remote sensing visual evidence reasoning," 2026. [Online]. Available: https://arxiv.org/abs/2605.17949
46. K. Ahn, S. Lee, K. P. Gummadi, and M. Cha, "GeoX: Mastering geospatial reasoning through self-play and verifiable rewards," 2026. [Online]. Available: https://arxiv.org/abs/2605.20006
47. M. Cai, G. Wang, W. Zhang, G. Zhou, Y. Zhuang, T. Zhang, H. Wang, H. Chen, and J. Li, "Earth-OneVision: Extending remote sensing multimodal large language models to more sensor modalities and tasks," 2026. [Online]. Available: https://arxiv.org/abs/2606.10819
48. D. T. Nguyen, T. Nguyen, F. A. Maani, H. M. Le, M. U. Sheikh, N. Saeed, M. H. Khan, and S. Khan, "TerraBench: Can agents reason over heterogeneous earth-system data?" 2026. [Online]. Available: https://arxiv.org/abs/2606.13148
49. X. He, T. Yang, T. Yan, H. Li, Y. Ge, Z. Ren, Z. Liu, J. Jiang, and C. Tang, "Efficient layer-wise cross-view calibration and aggregation for multispectral object detection," *Electronics*, vol. 15, no. 3, p. 498, Jan. 2026. [Online]. Available: https://doi.org/10.3390/electronics15030498
50. Y. Fu, X. Qiu, B. Ren, Y. Fu, R. Timofte, N. Sebe, M.-H. Yang, L. Van Gool, K. Zhang, Q. Nong *et al.*, "Ntire 2025 challenge on cross-domain few-shot object detection: Methods and results," in *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. IEEE, 2025, pp. 01–22.
51. X. Qiu, Y. Fu, J. Geng, B. Ren, J. Pan, Z. Wu, H. Tang, Y. Fu, R. Timofte, N. Sebe *et al.*, "The second challenge on cross-domain few-shot object detection at ntire 2026: Methods and results," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2026, pp. 2376–2405.
52. Y. Liu, J. Pan, J. Yang, T. Chen, P. Zhou, and B. Zhang, "Diverse instance generation via diffusion models for enhanced few-shot object detection in remote sensing images," *IEEE Geoscience and Remote Sensing Letters*, 2025.

 doi:10.20944/preprints202608.1840.v1

53. W. Li, X. Xiang, Z. Wen, G. Zhou, B. Niu, F. Wang, L. Huang, Q. Wang, and Y. Hu, "GeoReason: Aligning thinking and answering in remote sensing vision-language models via logical consistency reinforcement learning," *arXiv preprint arXiv:2601.04118*, 2026.
54. L. Sun, R. Fu, Z. Duan, H. Liu, X. Liu, and B. Yang, "GeoSolver: Scaling test-time reasoning in remote sensing with fine-grained process supervision," *arXiv preprint arXiv:2603.09551*, 2026.
55. X. He, H. Zhao, G. Wan, J. Pan, Y. Liu, X. Zou, Y. Luo, Y. Xu, J. Liu, W. Zhou, D. Tao, and B. Du, "Epistemic-aware vision-language foundation model for fetal ultrasound interpretation," in *Proceedings of the 32nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining, Volume 2*. ACM, Aug. 2026, pp. 11 061–11 072. [Online]. Available: https://doi.org/10.1145/3770855.3818860
56. G. Sumbul, A. de Wall, T. Kreuziger, F. Marcelino, H. Costa, P. Benevides, M. Caetano, B. Demir, and V. Markl, "BigEarthNet-MM: A large-scale, multimodal, multilabel benchmark archive for remote sensing image classification and retrieval [software and data sets]," *IEEE Geoscience and Remote Sensing Magazine*, vol. 9, no. 3, pp. 174–180, 2021. [Online]. Available: https://doi.org/10.1109/MGRS.2021.3089174
57. Z. Zhang, T. Zhao, Y. Guo, and J. Yin, "Rs5m and georsclip: A large scale vision-language dataset and a large vision-language model for remote sensing," *IEEE Transactions on Geoscience and Remote Sensing*, pp. 1–1, 2024.
58. Z. Wang, R. Prabha, T. Huang, J. Wu, and R. Rajagopal, "SkyScript: A large and semantically diverse vision-language dataset for remote sensing," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, no. 6, pp. 5805–5813, 2024. [Online]. Available: https://doi.org/10.1609/aaai.v38i6.28393
59. Z. Yuan, Z. Xiong, L. Mou, and X. X. Zhu, "ChatEarthNet: A global-scale image-text dataset empowering vision-language geo-foundation models," *Earth System Science Data*, vol. 17, no. 3, pp. 1245–1263, 2025. [Online]. Available: https://doi.org/10.5194/essd-17-1245-2025
60. Y. Zhao, M. Zhang, B. Yang, Z. Zhang, J. Kang, and J. Gong, "LuoJiaHOG: A hierarchy oriented geo-aware image caption dataset for remote sensing image-text retrieval," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 222, pp. 130–151, 2025. [Online]. Available: https://doi.org/10.1016/j.isprsjprs.2025.02.009
61. X. Li, J. Ding, and M. Elhoseiny, "VRSBench: A versatile vision-language benchmark dataset for remote sensing image understanding," in *Advances in Neural Information Processing Systems*, vol. 37, 2024, pp. 3229–3242. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2024/hash/05b7f821234f66b78f99e7803fffa78a-Abstract-Datasets_and_Benchmarks_Track.html
62. S. Ma, Z. Li, and J. A. Taylor, "Landsat30-AU: A vision-language dataset for australian landsat imagery," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 40, no. 10, pp. 7809–7817, 2026. [Online]. Available: https://doi.org/10.1609/aaai.v40i10.37724
63. A. Zavras, D. Michail, X. X. Zhu, B. Demir, and I. Papoutsis, "GAIA: A global, multimodal, multiscale vision-language dataset for remote sensing image analysis," *IEEE Geoscience and Remote Sensing Magazine*, vol. 14, no. 2, pp. 36–63, 2026. [Online]. Available: https://doi.org/10.1109/MGRS.2025.3650613
64. Y. He, X. Cheng, J. Zhu, C. Qiu, J. Wang, X. Zhang, Q. Huang, and K. Yang, "SAR-TEXT: A large-scale sar image-text dataset built with SAR-Narrator and a progressive learning strategy for downstream tasks," 2025. [Online]. Available: https://arxiv.org/abs/2507.18743
65. Y. Wei, A. Xiao, Y. Ren, Y. Zhu, H. Chen, J. Xia, and N. Yokoya, "SARLANG-1M: A benchmark for vision-language modeling in SAR image understanding," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 64, pp. 1–20, 2026. [Online]. Available: https://doi.org/10.1109/TGRS.2026.3652099
66. C. Liu, K. Chen, R. Zhao, Z. Zou, and Z. Shi, "Text2Earth: Unlocking text-driven remote sensing image generation with a global-scale dataset and a foundation model," *IEEE Geoscience and Remote Sensing Magazine*, vol. 13, no. 3, pp. 238–259, 2025. [Online]. Available: https://doi.org/10.1109/MGRS.2025.3560455
67. J. Pan, S. Lei, Y. Fu, J. Li, Y. Liu, Y. Sun, X. He, L. Peng, X. Huang, and B. Zhao, "Earthsynth: Generating informative earth observation with diffusion models," *arXiv preprint arXiv:2505.12108*, 2025.
68. J.-L. Herzog, M. J. Adler, L. Hackel, Y. Shu, A. Zavras, I. Papoutsis, P. Rota, and B. Demir, "BigEarthNet.txt: A large-scale multi-sensor image-text dataset and benchmark for earth observation," 2026. [Online]. Available: https://arxiv.org/abs/2603.29630
69. Y. Hu, J. Yuan, C. Wen, X. Lu, Y. Liu, and X. Li, "Rsgpt: A remote sensing vision language model and benchmark," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 224, pp. 272–286, 2025.
70. A. C. Karaca, E. Ozelbas, S. Berber, O. Karimli, T. Yildirim, and M. F. Amasyali, "Robust change captioning in remote sensing: SECOND-CC dataset and MModalCC framework," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 18, pp. 21 494–21 513, 2025. [Online]. Available: https://doi.org/10.1109/JSTARS.2025.3600613

 doi:10.20944/preprints202608.1840.v1

71. J. Wang, W. Xuan, H. Qi, Z. Liu, K. Liu, Y. Wu, H. Chen, J. Song, J. Xia, Z. Zheng, and N. Yokoya, "DisasterM3: A remote sensing vision-language dataset for disaster damage assessment and response," in *Advances in Neural Information Processing Systems*, vol. 38, 2025, pp. 179 475–179 489. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2025/hash/ec80d18205e39d42a27192d5f3ddd688-Abstract-Datasets_and_Benchmarks_Track.html
72. S. Tehrani, Y. Xu, L. Haglund, A. Berg, and M. Felsberg, "DisasterInsight: A multimodal benchmark for function-aware and grounded disaster assessment," 2026. [Online]. Available: https://arxiv.org/abs/2601.18493
73. X. Zhang, Z. Mai, Q. Li, Z. Liao, Y. Wen, Y. Chen, X. Fan, T. H. Chan, T. Bi, H. Liang, R. Su, Z. Qian, J. Zheng, J. Huang, Y. Lu, and H. Fu, "HM-Bench: A comprehensive benchmark for multimodal large language models in hyperspectral remote sensing," 2026. [Online]. Available: https://arxiv.org/abs/2604.08884
74. Z. Luo, D. Wang, H. Guo, J. Zhang, and B. Du, "VLRS-Bench: A vision-language reasoning benchmark for remote sensing," 2026. [Online]. Available: https://arxiv.org/abs/2602.07045
75. R. Fu, H. Liu, W. Zhang, Z. Lin, X. Yang, P. Zhang, and B. Yang, "OmniEarth: A benchmark for evaluating vision-language models in geospatial tasks," 2026. [Online]. Available: https://arxiv.org/abs/2603.09471
76. R. Ferrod, M. Lecene, K. Sapkota, G. Leifman, V. Silverman, G. Beryozkin, and S. Lobry, "GroundSet: A cadastral-grounded dataset for spatial understanding with vector data," 2026. [Online]. Available: https://arxiv.org/abs/2603.14609
77. M. Yang, Z. Zhou, S.-Y. Tian, K.-Y. Yu, L.-Z. Guo, and Y.-F. Li, "NeSy-Route: A neuro-symbolic benchmark for constrained route planning in remote sensing," 2026. [Online]. Available: https://arxiv.org/abs/2603.16307
78. R. Kazoom, Y. Gigi, G. Leifman, T. Shekel, and G. Beryozkin, "RSRCC: A remote sensing regional change comprehension benchmark constructed via retrieval-augmented best-of-N ranking," 2026. [Online]. Available: https://arxiv.org/abs/2604.20623
79. Z. Zhang, X. Zeng, X. Kong, K. Zhang, H. Liang, B. Shi, J. Zheng, J. Huang, Y. Lu, and H. Fu, "GTPBD-MM: A global terraced parcel and boundary dataset with multi-modality," 2026. [Online]. Available: https://arxiv.org/abs/2604.12315
80. F. Liu, D. Chen, Z. Guan, X. Zhou, J. Zhu, Q. Ye, L. Fu, and J. Zhou, "Remoteclip: A vision language foundation model for remote sensing," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 62, pp. 1–16, 2024. [Online]. Available: https://doi.org/10.1109/TGRS.2024.3390838
81. W. Chen, Y. Deng, J. Wei, J. Chen, J. Chen, Y. Feng, Z. Xi, D. Liu, K. Li, and Y. Meng, "Dgtrsd & dgtrs-clip: A dual-granularity remote sensing image-text dataset and vision language foundation model for alignment," 2025. [Online]. Available: https://arxiv.org/abs/2503.19311
82. J. Pan, Q. Ma, and C. Bai, "A prior instruction representation framework for remote sensing image-text retrieval," in *Proceedings of the 31st ACM International Conference on Multimedia*, 2023, pp. 611–620.
83. J. Pan, M. Ma, Q. Ma, C. Bai, and S. Chen, "Priorclip: Visual prior guided vision-language model for remote sensing image-text retrieval," *arXiv preprint arXiv:2405.10160*, 2024.
84. K. Klemmer, E. Rolf, C. Robinson, L. Mackey, and M. Rußwurm, "Satclip: Global, general-purpose location embeddings with satellite imagery," 2024. [Online]. Available: https://arxiv.org/abs/2311.17179
85. P. Jain, D. Marcos, D. Ienco, R. Interdonato, and T. Berchoux, "Timesenclip: A vision-language model for remote sensing using single-pixel time series," 2025. [Online]. Available: https://arxiv.org/abs/2508.11919
86. Y. Zhan, Z. Xiong, and Y. Yuan, "Skyeyegpt: Unifying remote sensing vision-language tasks via instruction tuning with large language model," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 221, pp. 64–77, 2025.
87. D. Muhtar, Z. Li, F. Gu, X. Zhang, and P. Xiao, "Lhrs-bot: Empowering remote sensing with vgi-enhanced large multimodal language model," in *Computer Vision – ECCV 2024*. Cham: Springer Nature Switzerland, 2025, pp. 440–457.
88. W. Zhang, M. Cai, T. Zhang, Y. Zhuang, J. Li, and X. Mao, "Earthmarker: A visual prompting multi-modal large language model for remote sensing," *IEEE Transactions on Geoscience and Remote Sensing*, 2024.
89. J. A. Irvin, E. R. Liu, J. C. Chen, I. Dormoy, J. Kim, S. Khanna, Z. Zheng, and S. Ermon, "Teochat: A large vision-language assistant for temporal earth observation data," in *International Conference on Learning Representations*, 2025.
90. P. Wang, H. Hu, B. Tong, Z. Zhang, F. Yao, Y. Feng, Z. Zhu, H. Chang, W. Diao, Q. Ye *et al.*, "Ringmogpt: A unified remote sensing foundation model for vision, language, and grounded tasks," *IEEE Transactions on Geoscience and Remote Sensing*, 2024.

 doi:10.20944/preprints202608.1840.v1

91. X. Liu and Z. Lian, "Rsunivlm: A unified vision language model for remote sensing via granularity-oriented mixture of experts," *arXiv preprint arXiv:2412.05679*, 2024.
92. Y. kelu, X. Nuo, Y. Rong, X. Yingying, G. Zhuoyan, K. Titinunt, R. yi, Z. Pu, W. Jin, W. Ning, and L. Chao, "Falcon: A remote sensing vision-language foundation model," *arXiv preprint arXiv:2503.11070*, 2025.
93. Y. Li, W. Guo, X. Yang, N. Liao, D. He, J. Zhou, and W. Yu, "Toward open vocabulary aerial object detection with clip-activated student-teacher learning," in *European Conference on Computer Vision*. Springer, 2024, pp. 431–448.
94. J. Kini, R. Gupta, and M. Shah, "Cross-view open-vocabulary object detection in aerial imagery," *arXiv preprint arXiv:2510.03858*, 2025.
95. Z. Huang, Y. Feng, Z. Liu, S. Yang, Q. Liu, and Y. Wang, "Openrsd: Towards open-prompts for object detection in remote sensing images," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2025, pp. 8384–8394.
96. J. Xie, G. Wang, T. Zhang, Y. Sun, H. Chen, Y. Zhuang, and J. Li, "Llama-unidetector: A llama-based universal framework for open-vocabulary object detection in remote sensing imagery," *IEEE Transactions on Geoscience and Remote Sensing*, 2025.
97. H. Hwang and S. S. Woo, "Fase: Feature-aligned scene encoding for open-vocabulary object detection in remote sensing," in *Proceedings of the 34th ACM International Conference on Information and Knowledge Management*, 2025, pp. 4822–4826.
98. Y. Zhou, M. Lan, X. Li, L. Feng, Y. Ke, X. Jiang, Q. Li, X. Yang, and W. Zhang, "Geoground: A unified large vision-language model for remote sensing visual grounding," *arXiv preprint arXiv:2411.11904*, 2024.
99. K. Li, D. Wang, T. Wang, F. Dong, Y. Zhang, L. Zhang, X. Wang, S. Li, and Q. Wang, "Rsvg-zeroov: Exploring a training-free framework for zero-shot open-vocabulary visual grounding in remote sensing images," *arXiv preprint arXiv:2509.18711*, 2025.
100. L. Yao, F. Liu, D. Chen, C. Zhang, Y. Wang, Z. Chen, W. Xu, S. Di, and Y. Zheng, "Remotesam: Towards segment anything for earth observation," in *Proceedings of the 33rd ACM International Conference on Multimedia*, 2025, pp. 3027–3036.
101. S. Ni, D. Wang, H. Chen, H. Guo, N. Zhang, and J. Zhang, "Unigeoseg: Towards unified open-world segmentation for geospatial scenes," *arXiv preprint arXiv:2511.23332*, 2025.
102. X. Ma, J. Li, C. Pei, and H. Liu, "Geomag: A vision-language model for pixel-level fine-grained remote sensing image parsing," in *Proceedings of the 33rd ACM International Conference on Multimedia*, 2025, pp. 5441–5450.
103. Y. Zheng, W. Wu, Q. Li, X. Wang, X. Zhou, A. Ren, J. Shen, L. Zhao, G. Li, and X. Yang, "Instruct-sam: A training-free framework for instruction-oriented remote sensing object recognition," *arXiv preprint arXiv:2505.15818*, 2025.
104. Y. Shu, B. Ren, Z. Xiong, D. Pani Paudel, L. Van Gool, B. Demir, N. Sebe, and P. Rota, "Earthmind: Towards multi-granular and multi-sensor earth observation with large multimodal models," *arXiv e-prints*, pp. arXiv–2506, 2025.
105. A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever, "Learning transferable visual models from natural language supervision," 2021. [Online]. Available: https://arxiv.org/abs/2103.00020
106. J. Pan, M. Ma, Q. Ma, C. Bai, and S. Chen, "Pir: Remote sensing image-text retrieval with prior instruction representation learning," *arXiv e-prints*, pp. arXiv–2405, 2024.
107. L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray, J. Schulman, J. Hilton, F. Kelton, L. Miller, M. Simens, A. Askell, P. Welinder, P. F. Christiano, J. Leike, and R. Lowe, "Training language models to follow instructions with human feedback," in *Advances in Neural Information Processing Systems*, vol. 35, 2022, pp. 27 730–27 744. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2022/file/b1efde53be364a73914f58805a001731-Paper-Conference.pdf
108. J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov, "Proximal policy optimization algorithms," 2017. [Online]. Available: https://arxiv.org/abs/1707.06347
109. R. Rafailov, A. Sharma, E. Mitchell, C. D. Manning, S. Ermon, and C. Finn, "Direct preference optimization: Your language model is secretly a reward model," in *Advances in Neural Information Processing Systems*, vol. 36, 2023, pp. 53 728–53 741. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2023/file/a85b405ed65c6477a4fe8302b5e06ce7-Paper-Conference.pdf

 doi:10.20944/preprints202608.1840.v1

110. Z. Shao, P. Wang, Q. Zhu, R. Xu, J. Song, X. Bi, H. Zhang, M. Zhang, Y. K. Li, Y. Wu, and D. Guo, "Deepseekmath: Pushing the limits of mathematical reasoning in open language models," 2024. [Online]. Available: https://arxiv.org/abs/2402.03300
111. Z. Liu, Z. Sun, Y. Zang, X. Dong, Y. Cao, H. Duan, D. Lin, and J. Wang, "Visual-RFT: Visual reinforcement fine-tuning," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2025, pp. 2034–2044. [Online]. Available: https://openaccess.thecvf.com/content/ICCV2025/html/Liu_Visual-RFT_Visual_Reinforcement_Fine-Tuning_ICCV_2025_paper.html
112. J. Zhang, J. Huang, H. Yao, S. Liu, X. Zhang, S. Lu, and D. Tao, "R1-VL: Learning to reason with multimodal large language models via step-wise group relative policy optimization," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2025, pp. 1859–1869. [Online]. Available: https://openaccess.thecvf.com/content/ICCV2025/html/Zhang_R1-VL_Learning_to_Reason_with_Multimodal_Large_Language_Models_via_ICCV_2025_paper.html
113. J. Wang, Y. Zhong, Z. Chen, Z. Zheng, A. Ma, and L. Zhang, "EarthVL: A progressive earth vision-language understanding and generation framework," *arXiv preprint arXiv:2601.02783*, 2026.
114. K. Li, Z. Xin, L. Pang, C. Pang, Y. Deng, J. Yao, G. Xia, D. Meng, Z. Wang, and X. Cao, "SegEarth-R1: Geospatial pixel reasoning via large language model," *arXiv preprint arXiv:2504.09644*, 2025.
115. Z. Xin, K. Li, L. Chen, W. Li, Y. Xiao, H. Qiao, W. Zhang, D. Meng, and X. Cao, "SegEarth-R2: Towards comprehensive language-guided segmentation for remote sensing images," *arXiv preprint arXiv:2512.20013*, 2025.
116. Y. Li, J. Pan, Z. Wei, J. Wang, M. Peng, and W. Xu, "GeoChrono: Benchmarking and rethinking long-term temporal understanding in remote sensing," 2026. [Online]. Available: https://arxiv.org/abs/2607.15768
117. M. Fiaz, H. Debary, P. Fraccaro, D. Paudel, L. Van Gool, F. Khan, and S. Khan, "GeoVLM-R1: Reinforcement fine-tuning for improved remote sensing reasoning," *arXiv preprint arXiv:2509.25026*, 2025. [Online]. Available: https://arxiv.org/abs/2509.25026
118. D. Wang, S. Liu, W. Jiang, F. Wang, Y. Liu, X. Qin, Z. Luo, C. Zhou, H. Guo, J. Zhang, B. Du, D. Tao, and L. Zhang, "GeoZero: Incentivizing reasoning from scratch on geospatial scenes," *arXiv preprint arXiv:2511.22645*, 2025. [Online]. Available: https://arxiv.org/abs/2511.22645
119. Z. Zhang, Z. Guan, T. Zhao, H. Shen, T. Li, Y. Cai, Z. Su, Z. Liu, J. Yin, and X. Li, "Geo-R1: Improving few-shot geospatial referring expression understanding with reinforcement fine-tuning," *arXiv preprint arXiv:2509.21976*, 2025.
120. A. Koksal and A. A. Alatan, "TinyRS-R1: Compact multimodal language model for remote sensing," *arXiv preprint arXiv:2505.12099*, 2025.
121. J. Liu, L. Sun, R. Fu, and B. 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. [Online]. Available: https://openreview.net/forum?id=lJ7zecny2e
122. R. Shao, Z. Li, Z. Zhang, L. Xu, X. He, H. Yuan, B. He, Y. Dai, Y. Yan, Y. Chen, W. Guo, and H. Li, "Asking like Socrates: Socrates helps VLMs understand remote sensing images," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2026, pp. 26 465–26 475. [Online]. Available: https://openaccess.thecvf.com/content/CVPR2026/html/Shao_Asking_like_Socrates_Socrates_helps_VLMs_understand_remote_sensing_images_CVPR_2026_paper.html
123. L. Yao, F. Liu, H. Lu, C. Zhang, R. Min, S. Xu, S. Di, and P. Peng, "RemoteReasoner: Towards unifying geospatial reasoning workflow," *arXiv preprint arXiv:2507.19280*, 2025.
124. K. Li, J. Wang, Z. Wang, H. Qiao, W. Zhang, D. Meng, and X. Cao, "Designing domain-specific agents via hierarchical task abstraction mechanism," *arXiv preprint arXiv:2511.17198*, 2025.
125. J. Sun, Y. Lin, Y. Xue, Y. Wang, Z. Yao, R. Qian, Z. Xu, J. Li, X. Liu, J. Pan *et al.*, "Pmmc: Prospective multimodal memory compilation for long-term lvlm agents," *arXiv preprint arXiv:2608.00962*, 2026.
126. C.-E. J. Yu, B. Jalaian, and N. D. Bastian, "Visual Reasoning Agent: Robust vision systems in remote sensing via inference-time scaling," 2025, accepted to MORS 2026 Artificial Intelligence Workshop Proceedings. [Online]. Available: https://arxiv.org/abs/2509.16343
127. A. Shabbir, M. A. Munir, A. Dudhane, M. U. Sheikh, M. H. Khan, P. Fraccaro, J. B. Moreno, F. S. Khan, and S. Khan, "ThinkGeo: Evaluating tool-augmented agents for remote sensing tasks," 2025. [Online]. Available: https://arxiv.org/abs/2505.23752
128. B. Wu, M. Fang, L. Chen, K. Xu, T. Cheng, and J. Wang, "Vision-language reasoning for geolocalization: A reinforcement learning approach," *arXiv preprint arXiv:2601.00388*, 2026.

 doi:10.20944/preprints202608.1840.v1

129. F. Schrodt, M. Beck, J. Estopinan, D. E. Bowler, C. Fontaine, P. Gaüzère, R. Goury, M. Grenié, I. S. Martins, N. Morueta-Holme *et al.*, "Advancing causal inference in ecology: Pathways for biodiversity change detection and attribution," *Methods in ecology and evolution*, vol. 16, no. 10, pp. 2276–2304, 2025.
130. X. Huang, J. Li, J. Pan, R. Debnath, and D. Guan, "Global hydropower infrastructure and its environmental risks mapped by multimodal ai," 2026.
131. M.-H. Nguyen, M.-P. T. Duong, Q.-L. Nguyen, V.-P. La, and V.-Q. Hoang, "In search of value: the intricate impacts of benefit perception, knowledge, and emotion about climate change on marine protection support," *Journal of Environmental Studies and Sciences*, vol. 15, no. 1, pp. 124–142, 2025.
132. Y. Sun, W. Luo, Y. Xiang, J. Pan, J. Li, Q. Zhang, and X. Huang, "Deploying atmospheric and oceanic ai models on chinese hardware and framework: Migration strategies, performance optimization and analysis," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 40, no. 47, 2026, pp. 40 521–40 527.
133. H. Lightman, V. Kosaraju, Y. Burda, H. Edwards, B. Baker, T. Lee, J. Leike, J. Schulman, I. Sutskever, and K. Cobbe, "Let's verify step by step," in *International Conference on Learning Representations*, 2024, pp. 39 578–39 601. [Online]. Available: https://proceedings.iclr.cc/paper_files/paper/2024/file/aca97732e30bcf1303bc22ac3924fd16-Paper-Conference.pdf
134. F. Wang, M. Chen, Y. Li, D. Wang, H. Wang, Z. Guo, Z. Wang, B. Shan, L. Lan, Y. Wang, H. Wang, W. Yang, B. Du, and J. Zhang, "Geollava-8k: Scaling remote-sensing multimodal large language models to 8k resolution," 2025. [Online]. Available: https://arxiv.org/abs/2505.21375
135. Y. Zhou, C. Jiang, C. Yuan, and J. Li, "Look where it matters: Training-free ultra-hr remote sensing vqa via adaptive zoom search," 2025. [Online]. Available: https://arxiv.org/abs/2511.20460
136. R. Liu, B. Fu, J. Song, K. Li, W. Li, L. Xue, H. Qiao, W. Zhang, D. Meng, and X. Cao, "Zoomearth: Active perception for ultra-high-resolution geospatial vision-language tasks," 2025. [Online]. Available: https://arxiv.org/abs/2511.12267
137. Y. Dang, M. Zhu, D. Wang, Y. Zhang, J. Yang, Q. Fan, Y. Yang, W. Li, F. Miao, and Y. Gao, "A benchmark for ultra-high-resolution remote sensing mllms," 2025. [Online]. Available: https://arxiv.org/abs/2512.17319