OpenEarthAgent: A Unified Framework for Tool-Augmented Geospatial Agents

Akashah Shabbir[1⋆], Muhammad Umer Sheikh[1⋆], Muhammad Akhtar Munir[1], Hiyam Debary[2], Mustansar Fiaz[2], Muhammad Zaigham Zaheer[1], Paolo Fraccaro[2], Fahad Shahbaz Khan[1,3], Muhammad Haris Khan[1], Xiao Xiang Zhu[4], and Salman Khan[1,5]

[1] Mohamed bin Zayed University of AI, United Arab Emirates
[2] IBM Research, Europe
[3] Linköping University, Sweden
[4] Technical University Munich, Germany
[5] Australian National University, Australia

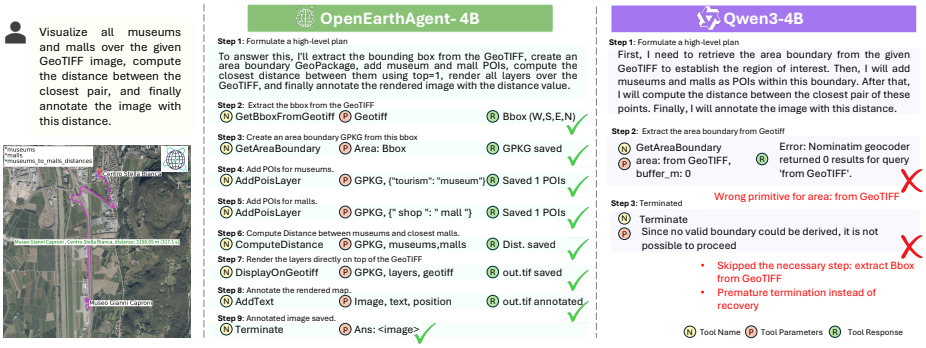


Fig. 1: Comparison of OpenEarthAgent-4B and Qwen3-4B on a complex GIS reasoning task. OpenEarthAgent correctly sequences tool calls with proper dependencies and feedback, while Qwen3 fails due to misordered tool usage and inconsistent reasoning.

Abstract. Recent progress in multimodal reasoning has enabled agents that interpret imagery, connect it with language, and execute structured analytical tasks. Extending these capabilities to remote sensing remains challenging, as models must reason over spatial scale, geographic structures, and multispectral indices while maintaining coherent multi-step logic. To address this gap, we introduce *OpenEarthAgent*, a unified framework for tool-augmented geospatial reasoning trained on satellite imagery, natural-language queries, and structured reasoning traces. Beyond serving as a benchmark, OpenEarthAgent establishes a cohesive agentic architecture built around a unified executable tool registry and trajectory-based policy learning. The framework standardizes heterogeneous visual, spectral, GIS, and georeferenced raster operations under

⋆ Equal contribution.

a consistent callable schema, enabling modular orchestration and deterministic execution. Training is performed via supervised fine-tuning on structured reasoning trajectories with deterministic replay validation to ensure executability and spatial correctness. The accompanying corpus comprises 14,538 training and 1,169 evaluation instances with over 107K reasoning steps, spanning urban, environmental, disaster, and infrastructure domains and incorporating GIS operations alongside index analyses such as NDVI, NBR, and NDBI. Grounded in explicit reasoning traces, the learned agent demonstrates structured reasoning, stable spatial understanding, and interpretable tool-driven behavior across diverse EO scenarios. We report consistent improvements over a strong baseline and competitive performance against recent open and closed-source models. `https://github.com/mbzuai-oryx/OpenEarthAgent`.

1 Introduction

The evolution of visual representation learning has transitioned from static perception to interactive multimodal reasoning. Early single-shot frameworks such as DINO [3] and MAE [13] laid the foundation for large-scale visual understanding through self-supervised objectives, learning strong image encoders without explicit language-based reasoning or interaction. Inspired by these successes, the remote sensing community extended this paradigm to earth observation, where models such as Prithvi [15], Copernicus-FM [40], Galileo [36], Panopticon [37], TerraFM [6], and AnySat [1] scaled representation learning to global multisensor data. These earth-scale vision models demonstrated remarkable transfer across spatial resolutions and modalities, yet their predictions remained one-shot and perception-centric, focusing on recognition rather than structured reasoning.

In parallel, large multi-modal models have progressed beyond perception. Architectures such as BLIP-2 [18], InstructBLIP [5], LLaVA-OneVision [17], and Kosmos-2 [27] extended language models into the visual domain, enabling grounded image understanding. Building on this trajectory, remote-sensing VLMs emerged to adapt multimodal reasoning to geospatial data. Early efforts such as RemoteCLIP [21], SkySenseGPT [23], and GeoChat [16] introduced large-scale multimodal alignment for remote sensing through paired image–text and instruction-following datasets. Recently, EarthDial [31] advanced this paradigm by extending grounding across optical, SAR, thermal, and temporal modalities, enabling unified multimodal reasoning for classification, captioning, and change analysis. Despite coupling language with perception, current remote sensing VLMs remain largely descriptive and lack explicit structured reasoning.

Subsequent reasoning-centric approaches such as ReAct [47], DeepSeek-R1 [10], and VLM-R1 [30] introduced structured planning and, in some cases, tool-driven reasoning, transforming static perception into organized multi-step task execution. Recent developments such as OpenThinkIMG [33] unified these advances through multi-step vision-tool interaction, where large vision-language models iteratively "*think with images*" through executable reasoning trajectories. Inspired by these advances, efforts such as ThinkGeo [29] and Earth-Agent [7]

began exploring tool-augmented reasoning for earth observation (EO). They demonstrate that large models can plan analytical chains but still face challenges in geospatial grounding, coordinate consistency, and physically verifiable outputs. These limitations highlight the need for geographically grounded agents that integrate perception with explicit and interpretable reasoning.

Building on this motivation, we construct a comprehensive agentic corpus and training framework that integrates GIS layers, index-driven computations, and multisensor imagery (optical and SAR) modalities to support a wide spectrum of EO reasoning tasks. Our dataset comprises training image-query reasoning-traces, along with a held-out evaluation set designed to benchmark model consistency and reasoning performance. The data span diverse contexts, including urban infrastructure, environmental monitoring, disaster assessment, land-use mapping, and transportation analysis, and incorporate GIS operations (e.g., distance, area, zonal statistics) alongside index-based analyses such as NDVI, NBR, and NDBI. Each sample provides an explicit reasoning trajectory linking multiple tool calls, intermediate states, and outcomes, enabling agents to learn structured, interpretable workflows rather than static predictions.

Central to this design is a unified tool registry that abstracts heterogeneous geospatial operators through a standardized executable schema, enabling consistent orchestration across perception, GIS computation, spectral analysis, and georeferenced raster reasoning. Unlike prior EO benchmarks that focus primarily on data scale, our contribution lies in coupling the dataset with an execution engine and trajectory-based policy alignment. This integration ensures that reasoning steps are not only generated but deterministically validated for spatial and logical consistency. To assess generalization and reasoning fidelity, we benchmark a broad suite of large and reasoning models, including the GPT family [25], Qwen [46], InternLM [2], and other recent multimodal LLMs. Our dataset is constructed through a unified data-gathering pipeline, where each modality undergoes question synthesis and automated validation to ensure reasoning diversity, spatial grounding, and cross-source consistency. This large-scale, tool-augmented corpus bridges the divide between GIS analysis and remote-sensing perception, providing a unified setting for evaluating structured geospatial reasoning. This work introduces an agentic framework for remote sensing that integrates tool-based reasoning within a unified training and evaluation setup.

Specifically, our key contributions are:

- **Unified data-construction pipeline:** A systematic process integrating multisensor imagery (RGB, SAR), GIS layers, and index-based computations through question synthesis and automated validation.
- **Agentic reasoning framework with deterministic validation:** A unified tool registry and a trajectory-based supervised alignment mechanism that enforces consistent tool syntax, executable reasoning chains, and spatially grounded decision-making (Fig. 1).
- **Comprehensive multimodal corpus:** 14,538 training instances and 1,169 evaluation tasks establishing a benchmark for studying spatial reasoning, grounding, and interpretability across reasoning and non-reasoning models.

2 Related Work

Remote sensing has progressed from task-specific models to large-scale foundation and vision-language models trained on global, multimodal data. Early self-supervised methods [4, 24] established transferable pretraining strategies on Sentinel archives. Subsequent foundation models include Prithvi-v2 [35] with spatiotemporal transformers on HLS, Copernicus-FM [40] with metadata-aware hypernetworks for Sentinel, Panopticon [37] and AnySat [1] with spectral- and resolution-adaptive embeddings, and CROMA [8] fusing radar and optical contrastive learning. Galileo [36] captures global-to-local context, demonstrating scalable pretraining across diverse sensors and geographies. On the multimodal side, GeoChat [16] enabled language grounding to satellite imagery, and EarthDial [31] extended this paradigm to multi-resolution optical, SAR, thermal, and temporal modalities across diverse tasks. Despite their scalability and representational strength, these models remain largely single-pass encoders; they lack structured reasoning, tool orchestration, and verifiable intermediate analysis.

Alongside foundation models, agentic systems have emerged that couple reasoning with external tools and feedback-driven control. Early frameworks such as ReAct [47] and Voyager [39] demonstrated the transition from static inference to autonomous, goal-directed reasoning. Subsequent systems, including WebAgent [42], VisTA [14], and DeepEyes [50], introduced modular architectures capable of selecting and sequencing tools under guided supervision, improving reasoning depth and execution consistency. OpenThinkIMG [33] unified these ideas through large-scale visual–tool integration with standardized APIs, while OctoTools [22] emphasized modular, verifiable execution via structured tool interfaces. Collectively, these systems established key principles: looped reasoning, standardized tool I/O, and trajectory-based learning, yet remain geospatially naive, lacking coordinate awareness, scale handling, and domain-specific spatial verification essential for earth observation analysis.

An emerging wave of EO research is now bringing these agentic principles into geospatial contexts. ThinkGeo [29] frames EO QA as tool-augmented reasoning, benchmarking ReAct-style chains and exposing persistent weaknesses in coordinate consistency, spatial grounding, and multi-step planning. Earth-Agent [7] broadens the tool ecosystem to include spectral products and standardized interfaces but still relies on largely predefined workflows, limiting physically verifiable GIS and index-based reasoning. Geo-OLM [32] explores plan-tool-verify prompting for compact models, improving efficiency but depending primarily on prompt-level heuristics rather than learned adaptive policies.

Building on this gap, we introduce an EO-native agentic stack that unifies dataset construction, supervised reasoning alignment, and evaluation under a physically grounded framework. OpenEarthAgent enables broad orchestration of GIS, index-based, optical, and SAR operations through structured instruction-tool pairs; trains on detailed reasoning traces with a dedicated evaluation set; and supports interpretable, verifiable, multi-step reasoning across diverse geospatial conditions.

Table 1: Comparison of geospatial/EO agent frameworks. Abbrev.: RS=remote sensing; MM=multimodal inputs; Tools=executable tool/function calls; MS=multi-step reasoning; GIS=GIS/vector-raster geospatial; Spec=spectral (e.g., indices, multi-spectral); CD=change detection; RData=reasoning-supervision data for training.

Method	RS	MM	MS	GIS	Spec	CD	RData
ReAct [47]	✗	✗	✓	✗	✗	✗	✗
OpenThinkIMG [33]	✗	✗	✓	✗	✗	✗	✓
ThinkGeo [29]	✓	✓	✓	✗	✗	✓	✗
Earth-Agent [7]	✓	✓	✓	✓	✓	✗	✗
Geo-OLM [32]	✓	✗	✓	✗	✗	✗	✗
RS-Agent [45]	✓	✗	✗	✗	✗	✗	✗
RS-ChatGPT [11]	✓	✗	✗	✗	✗	✗	✗
OpenEarthAgent (Ours)	✓	✓	✓	✓	✓	✓	✓

3 OpenEarthAgent

3.1 Dataset Curation

We curate a large-scale remote-sensing dataset linking imagery, queries, and tool-based reasoning traces to enable interpretable, multi-step geospatial analysis.

Overview and Motivation: The dataset presented in this work serves as a foundation for training and evaluating tool-augmented geospatial reasoning agents. While recent EO datasets primarily focus on visual classification or retrieval, they lack explicit reasoning structure and tool-level traceability. Our dataset bridges this gap by integrating imagery, natural-language queries, reasoning traces, and tool executions in a unified schema. It enables models to perform planning, verification, and computation beyond pattern recognition. A comparison against existing agent frameworks is provided in Table 1, highlighting differences in modality coverage, tool integration, reasoning depth, and supervision design. Consisting of 14,538 training samples and 1,169 held-out test samples for benchmarking, covering diverse geographic regions, sensor modalities, and reasoning tasks, it enables interpretable multi-step learning and evaluation.

Data Sources and Composition: The dataset integrates heterogeneous imagery and metadata from open-access EO repositories, encompassing optical, SAR, GIS, and multispectral domains. Primary sources include high-resolution benchmarks such as DOTA [43], DIOR [19], xBD [12], AID [44], NWPU-VHR-10 [38], FloodNet [28], and Global-Dumpsite [34], complemented by Sentinel collections and Google Earth Engine archives (see Table 2 for details). Each dataset offers distinct spatial resolutions and annotation formats (bounding boxes, segmentation masks, category labels), providing both fine-grained object understanding and large-scale semantic context. The collection spans seven thematic domains: *urban planning, disaster assessment, environmental monitoring, transportation, aviation, recreation, and industrial infrastructure*. Multispectral and indexed layers (e.g., NDVI, NBR, NDBI) are derived through Google Earth Engine to complement optical and SAR imagery with physically meaningful environmental indicators. Where required for spatial reasoning tasks, samples are

Table 2: Comprehensive overview of data sources used for dataset construction, across optical, SAR, GIS, and multispectral domains. These sources provide complementary resolutions and annotations spanning urban and transport monitoring to environmental and disaster assessment, forming a foundation for multi-sensor, tool-based geospatial reasoning. GSD: ground sample distance, B-Box: bounding box, Seg: segmentation.

Source	Annotation Type	Sensor (m/px)	Year	Applications / Tasks
Optical Datasets				
DIOR [19]	B-Box, Category	0.5–30	2024	Transport & Aviation Monitoring; Infrastructure Mapping
DOTA [43]	GSD, B-Box, Category	0.1–1	2021	Transport Surveillance; Infrastructure Assessment
NWPU-VHR-10 [38]	B-Box, Category	0.5–2	2023	Aviation; Transport; Infrastructure
UCAS-AOD [51]	B-Box, Category	0.5–2	2015	Aviation; Transport
AID [44]	B-Box, Category	0.2–2	2017	Urban Planning; Industrial Site Detection
iSAID [41]	GSD, B-Box, Seg. Map, Pix. Count	0.1–1	2019	Segmentation; Transport Monitoring
xBD [12]	GSD, B-Box, Category, Pix. Count	1–3.5	2019	Disaster Assessment; Change Detection
FloodNet [28]	GSD, B-Box, Category, Seg. Map, Pix. Count	0.015–0.02	2020	Flood Monitoring; Damage Estimation
Global-Dumpsite [34]	B-Box, Category	0.3–0.8	2023	Environmental Monitoring; Waste Localization
SAR Datasets				
SSDD [49]	B-Box, Category	1–15	2021	Ship Detection; Maritime Transport
SADD [48]	B-Box, Category	0.5–3	2022	Aviation; Vessel Identification
SIVED [20]	B-Box, Category	0.1–0.3	2023	Vehicle Detection; Transport Monitoring
GIS-based Sources				
OpenStreetMap [26]	POIs (Points of Interest)	N/A	2017-	Urban Mapping; Infrastructure Analysis
Indexed / Multispectral Sources				
GoogleEarthEngine [9]	Index Layers	10–30	2017-	Urban Planning; Change Detection; Environmental Monitoring

associated with geospatial metadata such as coordinates, projection reference, and ground sampling distance (GSD), enabling verifiable geospatial analysis.

Automated Curation Pipeline and Quality Control: Dataset construction begins by aggregating candidate regions, annotations, and metadata from multiple geospatial sources. For RGB-SAR branches, a *Sample Sufficiency Filter* module removes datapoints that do not meet the required annotation counts, after which an *Annotation Harmonizer* module aligns labels across source datasets to produce a unified annotation format. Index-based samples are mined by detecting temporal variations in spectral indicators such as vegetation loss, burn severity, or flooding. All sources are converted into a unified JSON schema encapsulating imagery paths, GeoPackage metadata, spatial attributes, and tool arguments. Before the final generation, candidate GIS queries are programmatically executed (e.g., POI counting, distance computation, area estimation) to

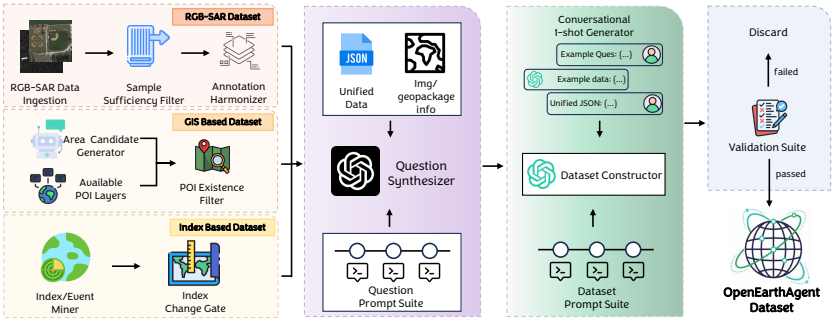


Fig. 2: Unified data-curation pipeline for the OpenEarthAgent dataset. RGB/SAR imagery, GIS spatial layers, and index-driven datasets are processed through dedicated filtering (annotation checks, POI constraints, and spectral-change signals such as NDVI/NBR/NDBI). The resulting samples are merged into unified JSON records with imagery metadata and geopackage information. A question synthesizer and conversational generator produce tool-grounded queries and reasoning traces, which are automatically validated to ensure geographic and syntactic correctness before forming the final corpus for agentic training and evaluation.

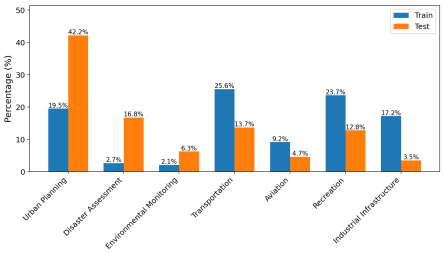
verify argument correctness and prevent formatting inconsistencies. Potential inconsistencies such as mismatches between query instructions and tool parameters or invalid execution outputs are automatically detected, followed by controlled regeneration and targeted manual verification when necessary.

Natural-language queries and reasoning trajectories are synthesized through a structured LLM-driven module, with task-specific prompt templates and employ one-shot exemplars tailored to each reasoning category. The data generator produces complete entries containing queries, multimodal inputs, and serialized reasoning traces. Each instance is validated through corresponding tool execution to verify argument correctness, geometric validity, and spatial consistency. All test-set samples undergo manual inspection to ensure realistic geospatial measurements (e.g., addressing GSD inconsistencies in area estimates). Validated samples form the final corpus. A summary of the pipeline is shown in Fig. 2.

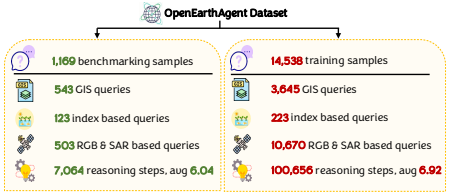
Task Structure and Query Design: Each instance represents a query-driven reasoning process that transforms static observation into an iterative perception-action loop (see Fig. 3). Queries are expressed in natural language and paired with structured reasoning traces that interleave *thoughts*, *actions*, and *observations*, ultimately converging on a final *answer*. These traces span diverse modalities and task types, including: GIS operations (distance, buffer, area, zonal statistics), object-level reasoning (counting, comparison, orientation, proximity), spectral index computation (NDVI, NBR, NDBI), and temporal change detection. Reasoning depth and tool sequence length vary across samples, but each trajectory provides a complete chain of intermediate steps, enabling transparent multi-step geospatial problem solving rather than one-shot prediction.

Dataset Statistics: Fig. 4 summarizes the scale, distribution, and structure of the OpenEarthAgent dataset. The training split contains 14,538 instances, and

Fig. 3: Representative query-reasoning trajectories from dataset, showing interleaving of natural-language thoughts, tool executions, and observations across tasks such as object localization, direction estimation, spectral mapping, change detection, and GIS-based analysis. Examples are *simplified excerpts* of longer multi-step reasoning traces.



(a) Data distribution across categories.



(b) Key statistics for the curated corpus used for training and evaluation of our proposed agentic pipeline.

Fig. 4: Dataset overview: category distribution (left) and corpus statistics (right).

the benchmark split contains 1,169 evaluation samples, each with multimodal imagery, a natural-language query, and a structured reasoning trace with interleaved tool calls and intermediate observations. The training set contains 100,656 reasoning steps (average 6.92 per query), and the benchmark split contains 7,064 reasoning steps (average 6.04 per query). Reasoning steps correspond to explicit thought-action-observation transitions, while tool calls capture the diversity of operational pathways required for solving curated agentic tasks. The breadth of modalities, thematic coverage, and depth of reasoning traces enable rigorous evaluation of grounded, interpretable, and tool-augmented geospatial reasoning.

3.2 Methodology

Overview: We present *OpenEarthAgent*, a unified agentic learning framework for multimodal reasoning over earth observation data. The framework integrates

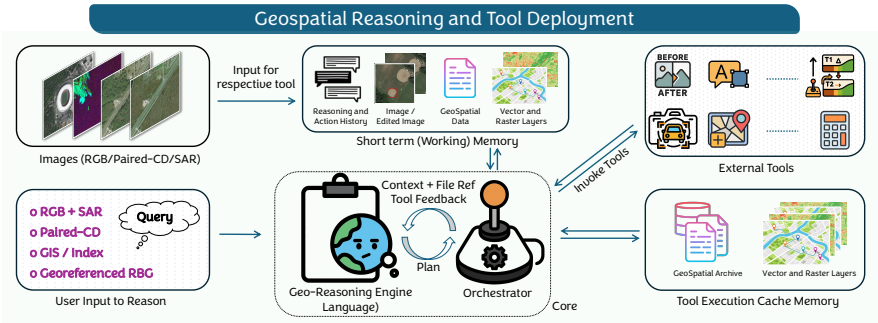


Fig. 5: Overview of the proposed *OpenEarthAgent* framework. The figure depicts tool deployment, where user queries over RGB, SAR, paired change-detection (CD), or GIS/indexed imagery are processed by the geospatial reasoning engine and tool orchestrator, which invoke appropriate tools and integrate feedback.

tool-driven interaction and language-grounded reasoning into a single end-to-end formulation. Unlike conventional, single-step multimodal models, OpenEarthAgent decomposes geospatial reasoning into an organized sequence of perception and action steps, where the agent learns to *perceive*, *reason*, and *act* through callable tools operating on visual, textual, and spatial modalities. Given a natural-language instruction $\mathbf{u}$ and corresponding multimodal inputs $\mathbf{v}$ (e.g., RGB imagery, paired change-detection inputs, SAR, GIS layers, or georeferenced raster data), the agent generates multi-step reasoning trajectories that interleave internal thoughts with explicit tool invocations and their returned outputs (see Fig. 5). At each step, the agent maintains a short-term working memory that contains: instructions, observations, prior tool calls, spatial metadata, and contextual feedback from previous executions, which enable iterative reasoning and consistent spatial grounding. Training follows a supervised fine-tuning procedure on validated reasoning traces, improving syntactic validity, spatial consistency, and multi-step reasoning coherence across diverse geospatial contexts.

Unified Tool Registry: To standardize the diverse operators used in EO reasoning, we construct a unified tool registry that abstracts heterogeneous visual, spectral, and GIS operations through a consistent callable schema:

$$\mathcal{M}_j = (\mathbf{x}_{\text{in}}, \mathbf{y}_{\text{out}}, \psi_j), \tag{1}$$

where $\mathbf{x}_{\text{in}}$ denotes structured input arguments, $\mathbf{y}_{\text{out}}$ denotes structured outputs, $\psi_j : \mathbf{x}_{\text{in}} \mapsto \mathbf{y}_{\text{out}}$ is the executable function implementing the operation. All tools follow standardized JSON-based contracts to ensure consistent serialization and validation across training and inference. A central orchestrator parses model-generated tool calls, validates arguments, executes ψ_j, and appends returned outputs to the working memory. During execution, a predicted tool invocation s_t with arguments $\hat{\mathbf{x}}_t$ is evaluated as $\mathbf{y}_t = \psi_{s_t}(\hat{\mathbf{x}}_t)$, and the returned output $\mathbf{y}_t$ becomes part of the updated reasoning context.

To ensure efficiency and consistency, intermediate outputs (e.g., derived vector layers, raster subsets, index maps, computed geometries) are stored in a tool execution cache memory. This cache includes geospatial archives and vector/raster layers that can be reused by subsequent tool calls within the same trajectory. Caching prevents redundant computation and ensures deterministic replay of reasoning steps. New tools can be integrated by registering their schema $\mathcal{M}_j$ without retraining the backbone model, ensuring extensibility while maintaining interface consistency. We outline the tool categories:
a) Perceptual Tools. ground language into image-space entities. `TextToBbox`, `ObjectDetection`, and `RegionAttributeDescription` convert free-form queries into spatial primitives such as bounding boxes and semantic tags. `CountGivenObject` and `SegmentObjectPixels` support object counting and pixel-level area estimation. `ChangeDetection` analyzes multi-date imagery to detect structural changes such as damage or flooding. **b) GIS Computation Tools.** enable geographic reasoning via geometric and coordinate-aware computation. `GetAreaBoundary`, `AddPoisLayer`, and `ComputeDistance` manipulate vector geometries within standard coordinate reference systems. Outputs include numerical measures (e.g., distances, buffers, zonal statistics) or derived layers visualized via `DisplayOnMap`. **c) Spectral Tools.** perform band arithmetic and spectral analysis on raster inputs. `AddIndexLayer`, `ComputeIndexChange`, and `ShowIndexLayer` compute and visualize indices (NDVI/NBR/NDBI), enabling quantitative reasoning about environmental dynamics. **d) Georeferenced Raster Tools.** `GetBboxFromGeotiff` and `DisplayOnGeotiff` operate directly on georeferenced rasters supporting coordinate extraction, bounding box derivation, and visualization aligned with spatial metadata, ensuring spatial consistency. **e) Utility Tools.** `Calculator`, `Solver`, and `Plot` support arithmetic and symbolic reasoning. Visualization tools `DrawBox` and `AddText` display results. `OCR` and `GoogleSearch` extend multimodal understanding through text extraction and contextual retrieval. `Terminate` signals completion of a reasoning trajectory.
Reasoning Trajectories and Training Objective: Each training sample encodes a reasoning trajectory that connects an instruction $\mathbf{u}$, observations $\mathbf{v}$, and a sequence of tool interactions. Formally, the trajectory is represented as:

$$\begin{aligned} &\Gamma_i = \{(s_t, r_t)\}_{t=1}^{T_i}, \quad s_t \in \mathcal{S}, \\ &r_t \sim \mathcal{E}\big(s_t \mid \mathbf{u}_i, \mathbf{v}_i, \{(s_k, r_k)\}_{k<t}\big), \end{aligned} \tag{2}$$

where $\mathcal{S}$ denotes the set of registered tools, $\mathcal{E}$ the execution environment (tool controller + cache), and (s_t, r_t) is the t-th action-result pair. Each s_t is a predicted tool invocation, while r_t is the observation returned by the environment.

The autoregressive model conditions on the full working memory state when predicting the next action. Training is performed via maximum likelihood estimation over verified trajectories, optimizing only the tool-action policy:

$$\mathcal{L}_{\text{train}} = -\frac{1}{N} \sum_{i=1}^{N} \sum_{t=1}^{T_i} \log P_\eta(s_t \mid \mathbf{u}_i, \mathbf{v}_i, s_{<t}, r_{<t}), \tag{3}$$

where η parameterizes the model and N is the number of training samples. Tool observations r_t are added to the working memory as context but are masked from loss computation and not treated as policy outputs. This formulation enforces syntactic validity of tool calls, sequential and spatial coherence, and argument correctness, while preserving the environment-policy separation.

Before inclusion in training, each trajectory undergoes deterministic replay through the tool controller. Re-execution verifies argument formatting, coordinate integrity, geometric validity, and full-chain executability. Through supervised fine-tuning on validated multi-step trajectories, OpenEarthAgent learns to generate syntactically correct, spatially grounded, and interpretable reasoning workflows over heterogeneous EO data.

4 Experiments

In this section, we outline implementation details, describe the two evaluation modes, define metrics used to quantify reasoning fidelity, tool effectiveness, and geospatial accuracy, and discuss key experimental results and insights.

Implementation Details: OpenEarthAgent is finetuned using Qwen3-4B-Instruct-2507 as the base model. Training was performed on LLM controller on four NVIDIA A100 GPUs (40 GB). The base model is loaded via Unsloth's FastLanguageModel, enabling efficient multi-GPU distributed training. The model is trained for one epoch with a learning rate of 2×10^{-5}, cosine learning rate scheduling, 0.05 warmup ratio, and a batch size of 16. The maximum sequence length is set to 4096 tokens to support long-context conversational tool interactions. LLM is finetuned on conventional data, where the model alternates between text generation and tool invocation, providing tool outputs and feedback in JSON format, embedded in its user and assistant turns. Training applies response-only masking, computing loss exclusively on assistant-generated tokens, enabling accurate tool calls and structured responses without optimizing on prompt text or external tool outputs. This enables the model to establish a good grounding for tool invocation.

4.1 Baseline Evaluation

To evaluate OpenEarthAgent, we benchmark its performance against a diverse set of large language models (LLMs), encompassing both proprietary and open-source variants. The comparison includes advanced frontier models (GPT-4o and its reasoning-optimized variant o4-mini), and mid- to small-scale open-weight models, including Llama3.1 and Internlm3. All models serve as interchangeable backbones within the unified OpenEarthAgent framework. We assess performance using both step-by-step and end-to-end protocols, allowing analysis of intermediate reasoning trajectories and final task completion accuracy.

Step-by-Step Evaluation: This mode measures procedural reasoning and tool-invocation behavior without executing tools. The model is given n reasoning steps and must generate a valid action at each one based on the full interaction

history. The first step is exempt from validation to allow the model to formulate an initial high-level plan before committing to concrete tool interactions. This setup isolates reasoning quality and enables focused analysis of spatial semantics, plan coherence, and geospatial context understanding.

End-to-End Evaluation: This mode tests full autonomous execution with live tool use. The model issues tool calls, forms arguments, and reasons step-by-step based on prior outputs. This captures operational robustness, error handling, and the effectiveness of linking perception to action in geospatial workflows.

Evaluation Metrics: We assess the model's procedural reasoning with five metrics. Instance Accuracy (Inst.) measures the proportion of tool called without logical or syntactic errors; Tool Accuracy (Tool) evaluates correctness of tool selection; Argument Name Accuracy (ArgN) checks the presence of all required arguments in the generated action; Argument Value Accuracy (ArgV) verifies argument correctness; and Summarization Accuracy (SummAcc) quantifies how well the model consolidates information from prior tool calls into a coherent final response. We compute F_1 for tool selection across four functional categories: Perception (Per.), Operation (Op.), Logic (Logic.), and GIS (GIS.) to capture alignment between predicted and reference tool sets. Trajectory quality is further assessed via tool-order accuracy under increasing constraints: *Unique* checks tool-to-tool correspondence without penalizing repeated calls, *AnyOrder* requires all tools (including duplicates) regardless of order, and *SameOrder* enforces both correct sequence and frequency of tool invocations. Overall task Accuracy compares final model completions with reference outputs. For image-generation tasks, success is determined by the correctness of generation tool calls and their execution success rate; all non-generation queries are evaluated by an LLM judge (gpt-4o-mini) using a standardized elaboration evaluation prompt.

Table 3: Step-by-step evaluation under tool-agnostic rollouts. We report accuracies: Inst.; syntactic + logical validity of actions, Tool.; correct tool selection, ArgN.; presence of required parameters, ArgV.; parameter correctness, and Summ.; correctness of the final consolidated answer. While frontier LLMs lead on summary accuracy, OpenEarthAgent attains state-of-the-art Inst./Tool./ArgN./ArgV., narrowing the gap despite a substantially smaller parameter budget.

Model	Param.	Inst.	Tool.	ArgN.	ArgV.	Summ.
FrontierLLMs						
gpt-5		97.54	82.69	73.28	37.28	87.02
gpt-4o		99.18	93.88	85.48	45.80	86.76
o4-mini		83.68	68.33	64.17	37.95	**89.48**
Open-SourceLLMs						
Qwen2.5-Instruct	7B	94.08	85.51	78.46	38.00	80.08
Llama-3.1-Instruct	8B	47.07	39.30	34.11	17.49	79.08
Internlm3-Instruct	8B	44.54	38.46	27.82	13.25	29.84
Mistralv0.3-Instruct	7B	64.14	35.12	26.11	12.45	24.04
Qwen2.5-Instruct	3B	85.07	72.13	64.87	24.12	68.75
BaselineLLM						
Qwen3-Instruct-2507	4B	97.34	86.94	84.12	33.55	83.28
OpenEarthAgent						
OpenEarthAgent	4B	**99.51**	**97.18**	**96.08**	**62.10**	83.64

Table 4: End-to-end evaluation: (i) F_1 scores for tool selection across: Perception (Per.), Operation (Op.), Logic (Logic.), and GIS (GIS.); (ii) tool-order fidelity: *AnyOrder* (multiset match, order-agnostic), *SameOrder* (exact sequence match), and *Unique* (set match, no multiplicity); and (iii) task-level Accuracy, *Ans.* for non-generative and *Gen.* for image-generation. GPT-4o attains high GIS F_1, while OpenEarthAgent achieves strong, balanced performance with superior trajectory fidelity.

Model	Param.	F_1 scores				Tool Order			Accuracy	
		Per.	Op.	Logic.	GIS.	AnyOr.	SameO.	Uni.	Ans.	Gen.
FrontierLLMs										
gpt-5		16.88	47.00	6.26	91.50	46.96	46.79	47.81	43.88	46.21
gpt-4o		44.47	**66.91**	35.95	95.80	50.81	50.38	55.52	39.22	**77.93**
o4-mini		39.48	64.93	12.38	86.51	40.12	39.95	41.49	35.18	55.17
Open-SourceLLMs										
Qwen2.5-Instruct	7B	20.38	37.33	26.68	75.46	31.57	30.02	36.61	15.85	41.38
Llama-3.1-Instruct	8B	22.54	16.45	32.03	64.64	39.01	37.72	44.91	12.70	55.17
Internlm3-instruct	8B	5.50	14.21	0.51	31.47	2.22	1.88	3.16	0.24	2.76
Mistral-Instruct-v0.3	7B	6.74	10.24	15.82	17.12	2.65	2.57	2.91	0.39	5.52
Qwen2.5-Instruct	3B	15.63	16.54	11.16	35.64	9.24	8.72	12.40	1.83	24.14
BaselineLLM										
Qwen3-Instruct-2507	4B	13.86	20.95	18.35	71.82	16.00	14.71	21.47	13.72	15.86
OpenEarthAgent										
OpenEarthAgent	4B	**58.30**	56.76	**51.18**	**98.52**	**67.75**	**67.24**	**72.71**	**45.26**	75.86

4.2 Results

In step-by-step evaluation (Table 3), frontier models lead: o4-mini score highest on Summ. (89.48) while gpt-4o shows strong Inst. Acc. Among open-source models, Qwen2.5-I (7B) performs competitively (Inst. 94.08, Tool. 85.51), whereas other 7B–8B models degrade, particularly on ArgV. prediction. OpenEarthAgent (4B) substantially narrows this gap despite its smaller size, achieving state-of-the-art Inst./ Tool./ ArgN./ and ArgV.(99.51/97.18/96.08/62.10). In the end-to-end evaluation (Table 4), GPT-5 avoids external logic tools and relies on internal numerical and analytical computations. GPT-4o leads in Op. F_1 (66.91) and Gen. Acc. (77.93) but remains modest in Per./Logic, whereas OpenEarthAgent delivers balanced gains across Per./Op./Logic./GIS and dominates in tool-order accuracy (AnyOr./SameO. /Uni. $\approx$ 67-72%), while substantially improving Answer Acc. (45.26) indicating robust trajectory planning rather than isolated tool correctness. Overall, these results suggest that targeted training with tool schemas and trajectories can outperform larger general-purpose models for geospatial tool use, especially under strict order/frequency constraints.

Cross-Benchmark Evaluation We extend our evaluation to Earth-Agent [7] benchmark(248 tasks, 13,729 images spanning the spectrum, products, and RGB modalities), scoring with fine-grained step-level metrics. For compatibility, benchmark tools are reformatted to the OpenEarthAgent interface, where curated JSON exemplars standardize tool invocation. As Table 5 reports, OpenEarthAgent (4B) outperforms similarly sized open-source baselines on every tool-related metric, and rivals GPT-4o, showing that structured agent design and training bring small models to within reach of frontier-level step-wise geospatial reasoning and tool execution. These gains transfer to ThinkGeo, evident from Table

Table 5: Step-by-step evaluation on Earth-Agent Benchmark.

Model	Param.	Inst.	Tool.	ArgN.	ArgV.	Summ.
FrontierLLMs						
gpt-4o		**97.41**	**71.65**	**70.01**	**48.11**	**80.26**
Open-SourceLLMs						
Qwen2.5-Instruct	7B	85.98	57.76	56.06	34.66	79.60
Llama-3.1-Instruct	8B	83.21	55.11	54.29	35.98	73.05
BaselineLLM						
Qwen3-Instruct-2507	4B	88.32	60.61	59.53	38.89	76.79
OpenEarthAgent						
OpenEarthAgent	4B	96.91	70.33	69.13	44.44	79.33

Table 6: (Left) ThinkGeo benchmark comparison on answer-level metrics. OpenEarthAgent achieves the scores, outperforming ThinkGeo and Qwen3-4B. (Right) Comparison with a generic GPT agent across the spectrum, product, and image tasks. OpenEarthAgent achieves the best overall performance with the lowest latency.

Model	Ans	Ans_I
ThinkGeo (reported)	9.78	20.40
Qwen3-4B (BL)	12.53	25.92
OpenEarthAgent	**16.35**	**29.13**

Model	Spec.	Prod.	Img.	Total	Lat(s)
GPT-Agent	14.67	22.25	28.33	21.75	933.54
Qwen3-4B (BL)	31.07	13.98	23.66	22.90	42.02
OpenEarthAgent	**90.46**	**60.67**	**29.00**	**60.04**	**21.67**

6 (Left), OpenEarthAgent attains the highest Ans (16.35) and Ans_I (29.13), surpassing both the reported ThinkGeo results and the Qwen3-4B baseline. Such consistency across benchmarks confirms that OpenEarthAgent couples reasoning and tool use into a robust, transferable capability.

Comparison with General Agent To assess OpenEarthAgent against a general purpose agent, we evaluate a balanced 60-query subset(20 spectrum, 20 product, and 20 image-analysis tasks) spanning the benchmark's full range of reasoning and tool-use demands. Table 6 (Right) shows OpenEarthAgent leading in every category, with the widest margins on spectrum tasks, where it scores 90.46, more than 4 times GPT-Agent's 14.67. Overall, OpenEarthAgent achieves 60.04 while `GPT-Agent` reaches only 21.75, and the `Qwen3-4B` baseline reaches 22.90. OpenEarthAgent is also far more efficient, averaging just 21.67 seconds per query versus 42.02 seconds for Qwen3-4B ($\tilde{2}$x slower) and 933.54 seconds for GPT-Agent ($\tilde{4}$4x slower). OpenEarthAgent's task-specific tool selection yields higher accuracy and lower latency than generic problem-solving.

Tool Call & Success Patterns: Proprietary models (gpt-5, gpt-4o and o4-mini) show strong tool-control, issuing thousands of calls with low error rates ($\sim$7%), indicating stable formatting. OpenEarthAgent executes a higher volume of tool calls while achieving a superior overall success rate. In contrast, open-source models such as Qwen2.5-7B, Qwen2.5-3B, and Llama-3.1-8B exhibit high failure rates (43-46%), reflecting weak adherence to the tool schema. Mistral-7B demonstrates a lower error rate but fails to invoke tools at the required adequacy to meet task demands. Despite extensive tool usage, Llama-3.1-8B and Mistral-7B fail to converge to a final answer within the allowed step, resulting in 94%

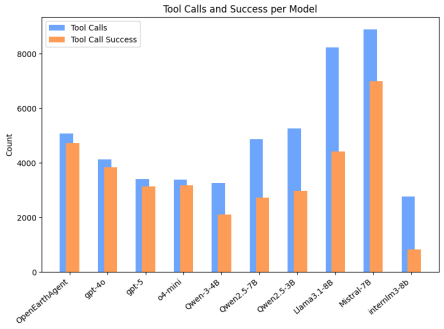


Fig. 6: Tool-call performance: Open-source models show lower success rates, while GPT family and OpenEarthAgent achieve substantially higher success rates.

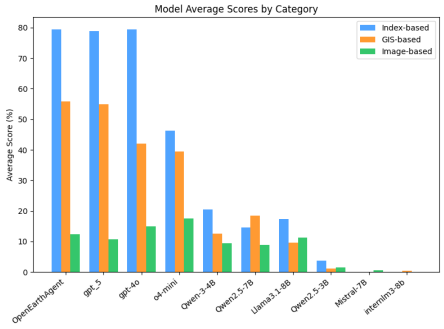


Fig. 7: Category-wise performance: OpenEarthAgent and gpt-5 lead in index/GIS score, while open source performs comparatively on image-based queries.

and 98% incomplete task executions, respectively. Overall, performance as shown in Fig. 6 varies by call volume and precision: robust models maintain low error rates, while others struggle with consistent, well-formed tool invocations.

Category Scores: As seen in Fig. 7, OpenEarthAgent (79.43%), gpt-4o (79.39 %), and gpt-5 (78.94%) achieve the strongest index-based scores, substantially outperforming all baselines. GIS-based tasks reveal a sharper degradation, with OpenEarthAgent (55.77 %) followed by gpt-5 (54.94 %), gpt-4o (41.95 %), and o4-mini (39.55%). Open-source models; qwen2.5-7b, qwen3-4b, and llama3.1-8b perform comparably well on image-based queries, while smaller models approach zero on index tasks and score only marginally on GIS and image-based tasks. Overall, the results highlight a significant performance gap and show that index- and GIS-based reasoning remains difficult for smaller open models. Additional analyses and ablations are provided in the supplementary material.

5 Conclusion

We introduced *OpenEarthAgent*, a unified framework for tool-augmented geospatial reasoning that connects natural language with multimodal earth observation data through structured analytical workflows. Unlike conventional EO models that focus primarily on perception, OpenEarthAgent performs grounded and interpretable reasoning by orchestrating visual, spectral, GIS, and GeoTIFF-aware tools under a consistent executable schema. The accompanying corpus comprises 14,538 training instances and 1,169 held-out test instances with validated multi-step reasoning trajectories, including explicit tool calls, intermediate observations, and outputs. OpenEarthAgent addresses a key gap in EO research by moving beyond perception-only modeling toward structured and verifiable geospatial reasoning. We believe this work provides a step toward agentic AI systems for environmental monitoring, disaster response, infrastructure analysis, and broader geospatial decision-making.

References

1. Astruc, G., Gonthier, N., Mallet, C., Landrieu, L.: Anysat: One earth observation model for many resolutions, scales, and modalities. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 19530–19540 (2025)
2. Cai, Z., Cao, M., Chen, H., Chen, K., Chen, K., Chen, X., Chen, X., Chen, Z., Chen, Z., Chu, P., Dong, X., Duan, H., Fan, Q., Fei, Z., Gao, Y., Ge, J., Gu, C., Gu, Y., Gui, T., Guo, A., Guo, Q., He, C., Hu, Y., Huang, T., Jiang, T., Jiao, P., Jin, Z., Lei, Z., Li, J., Li, J., Li, L., Li, S., Li, W., Li, Y., Liu, H., Liu, J., Hong, J., Liu, K., Liu, K., Liu, X., Lv, C., Lv, H., Lv, K., Ma, L., Ma, R., Ma, Z., Ning, W., Ouyang, L., Qiu, J., Qu, Y., Shang, F., Shao, Y., Song, D., Song, Z., Sui, Z., Sun, P., Sun, Y., Tang, H., Wang, B., Wang, G., Wang, J., Wang, J., Wang, R., Wang, Y., Wang, Z., Wei, X., Weng, Q., Wu, F., Xiong, Y., Xu, C., Xu, R., Yan, H., Yan, Y., Yang, X., Ye, H., Ying, H., Yu, J., Yu, J., Zang, Y., Zhang, C., Zhang, L., Zhang, P., Zhang, P., Zhang, R., Zhang, S., Zhang, S., Zhang, W., Zhang, W., Zhang, X., Zhang, X., Zhao, H., Zhao, Q., Zhao, X., Zhou, F., Zhou, Z., Zhuo, J., Zou, Y., Qiu, X., Qiao, Y., Lin, D.: Internlm2 technical report (2024)
3. Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., Joulin, A.: Emerging properties in self-supervised vision transformers. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 9650–9660 (2021)
4. Cong, Y., Khanna, S., Meng, C., Liu, P., Rozi, E., He, Y., Burke, M., Lobell, D., Ermon, S.: Satmae: Pre-training transformers for temporal and multi-spectral satellite imagery. Advances in Neural Information Processing Systems **35**, 197–211 (2022)
5. Dai, W., Li, J., Li, D., Tiong, A., Zhao, J., Wang, W., Li, B., Fung, P.N., Hoi, S.: Instructblip: Towards general-purpose vision-language models with instruction tuning. Advances in neural information processing systems **36**, 49250–49267 (2023)
6. Danish, M.S., Munir, M.A., Shah, S.R.A., Khan, M.H., Anwer, R.M., Laaksonen, J., Khan, F.S., Khan, S.: TerraFM: A scalable foundation model for unified multi-sensor earth observation. In: The Fourteenth International Conference on Learning Representations (2026)
7. Feng, P., Lv, Z., Ye, J., Wang, X., Huo, X., Yu, J., Xu, W., Zhang, W., Bai, L., He, C., Li, W.: Earth-agent: Unlocking the full landscape of earth observation with agents. In: International Conference on Learning Representations (ICLR) (2026)
8. Fuller, A., Millard, K., Green, J.: Croma: Remote sensing representations with contrastive radar-optical masked autoencoders. Advances in Neural Information Processing Systems **36**, 5506–5538 (2023)
9. Google Earth Engine: Google earth engine. `https://earthengine.google.com/` (2025), accessed: 29 June 2026
10. Guo, D., Yang, D., Zhang, H., Song, J., Wang, P., Zhu, Q., Xu, R., Zhang, R., Ma, S., Bi, X., et al.: Deepseek-r1 incentivizes reasoning in llms through reinforcement learning. Nature **645**(8081), 633–638 (2025)
11. Guo, H., Su, X., Wu, C., Du, B., Zhang, L., Li, D.: Remote sensing chatgpt: Solving remote sensing tasks with chatgpt and visual models. In: IGARSS 2024-2024 IEEE International Geoscience and Remote Sensing Symposium. pp. 11474–11478. IEEE (2024)
12. Gupta, R., Hosfelt, R., Sajeev, S., Patel, N., Goodman, B., Doshi, J., Heim, E., Choset, H., Gaston, M.: xbd: A dataset for assessing building damage from satellite imagery. arXiv preprint arXiv:1911.09296 (2019)

13. He, K., Chen, X., Xie, S., Li, Y., Dollár, P., Girshick, R.: Masked autoencoders are scalable vision learners. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 16000–16009 (2022)
14. Huang, Z., Ji, Y., Rajan, A.S., Cai, Z., Xiao, W., Wang, H., Hu, J., Lee, Y.J.: VisualToolAgent (VisTA): A reinforcement learning framework for visual tool selection. arXiv preprint arXiv:2505.20289 (2025)
15. Jakubik, J., Roy, S., Phillips, C., Fraccaro, P., Godwin, D., Zadrozny, B., Szwarcman, D., Gomes, C., Nyirjesy, G., Edwards, B., et al.: Foundation models for generalist geospatial artificial intelligence. arXiv preprint arXiv:2310.18660 (2023)
16. Kuckreja, K., Danish, M.S., Naseer, M., Das, A., Khan, S., Khan, F.S.: Geochat: Grounded large vision-language model for remote sensing. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 27831–27840 (2024)
17. Li, B., Zhang, Y., Guo, D., Zhang, R., Li, F., Zhang, H., Zhang, K., Zhang, P., Li, Y., Liu, Z., et al.: Llava-onevision: Easy visual task transfer. arXiv preprint arXiv:2408.03326 (2024)
18. Li, J., Li, D., Savarese, S., Hoi, S.: Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In: International conference on machine learning. pp. 19730–19742. PMLR (2023)
19. Li, K., Wan, G., Cheng, G., Meng, L., Han, J.: Object detection in optical remote sensing images: A survey and a new benchmark. ISPRS journal of photogrammetry and remote sensing **159**, 296–307 (2020)
20. Lin, X., Zhang, B., Wu, F., Wang, C., Yang, Y., Chen, H.: Sived: A sar image dataset for vehicle detection based on rotatable bounding box. Remote Sensing **15**(11), 2825 (2023)
21. Liu, F., Chen, D., Guan, Z., Zhou, X., Zhu, J., Ye, Q., Fu, L., Zhou, J.: Remoteclip: A vision language foundation model for remote sensing. IEEE Transactions on Geoscience and Remote Sensing **62**, 1–16 (2024)
22. Lu, P., Chen, B., Liu, S., Thapa, R., Boen, J., Zou, J.: Octotools: A multi-agent framework with extensible tools for complex reasoning. In: Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) (2026)
23. Luo, J., Pang, Z., Zhang, Y., Wang, T., Wang, L., Dang, B., Lao, J., Wang, J., Chen, J., Tan, Y., et al.: Skysensegpt: A fine-grained instruction tuning dataset and model for remote sensing vision-language understanding. arXiv preprint arXiv:2406.10100 (2024)
24. Manas, O., Lacoste, A., Giró-i Nieto, X., Vazquez, D., Rodriguez, P.: Seasonal contrast: Unsupervised pre-training from uncurated remote sensing data. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 9414–9423 (2021)
25. OpenAI: Introducing gpt. `https://openai.com` (2025), accessed: 29 June 2026
26. OpenStreetMap contributors: Openstreetmap: Collaborative project to create a free editable map of the world. `https://www.openstreetmap.org` (2025), accessed: 29 June 2026
27. Peng, Z., Wang, W., Dong, L., Hao, Y., Huang, S., Ma, S., Wei, F.: Kosmos-2: Grounding multimodal large language models to the world. arXiv preprint arXiv:2306.14824 (2023)
28. Rahnemoonfar, M., Chowdhury, T., Sarkar, A., Varshney, D., Yari, M., Murphy, R.R.: Floodnet: A high resolution aerial imagery dataset for post flood scene understanding. IEEE Access **9**, 89644–89654 (2021)

29. Shabbir, A., Munir, M.A., Dudhane, A., Sheikh, M.U., Khan, M.H., Fraccaro, P., Moreno, J.B., Khan, F.S., Khan, S.: Thinkgeo: Evaluating tool-augmented agents for remote sensing tasks. arXiv preprint arXiv:2505.23752 (2025)
30. Shen, H., Liu, P., Li, J., Fang, C., Ma, Y., Liao, J., Shen, Q., Zhang, Z., Zhao, K., Zhang, Q., et al.: Vlm-r1: A stable and generalizable r1-style large vision-language model. arXiv preprint arXiv:2504.07615 (2025)
31. Soni, S., Dudhane, A., Debary, H., Fiaz, M., Munir, M.A., Danish, M.S., Fraccaro, P., Watson, C.D., Klein, L.J., Khan, F.S., et al.: Earthdial: Turning multi-sensory earth observations to interactive dialogues. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 14303–14313 (2025)
32. Stamoulis, D., Marculescu, D.: Geo-olm: Enabling sustainable earth observation studies with cost-efficient open language models & state-driven workflows. In: Proceedings of the ACM SIGCAS/SIGCHI Conference on Computing and Sustainable Societies. pp. 608–619 (2025)
33. Su, Z., Li, L., Song, M., Hao, Y., Yang, Z., Zhang, J., Chen, G., Gu, J., Li, J., Qu, X., et al.: Openthinkimg: Learning to think with images via visual tool reinforcement learning. arXiv preprint arXiv:2505.08617 (2025)
34. Sun, X., Yin, D., Qin, F., Yu, H., Lu, W., Yao, F., He, Q., Huang, X., Yan, Z., Wang, P., et al.: Revealing influencing factors on global waste distribution via deep-learning based dumpsite detection from satellite imagery. Nature Communications **14**(1), 1444 (2023)
35. Szwarcman, D., Roy, S., Fraccaro, P., Gíslason, Þ.E., Blumenstiel, B., Ghosal, R., de Oliveira, P.H., Almeida, J.L.d.S., Sedona, R., Kang, Y., et al.: Prithvi-eo-2.0: A versatile multi-temporal foundation model for earth observation applications. arXiv preprint arXiv:2412.02732 (2024)
36. Tseng, G., Fuller, A., Reil, M., Herzog, H., Beukema, P., Bastani, F., Green, J.R., Shelhamer, E., Kerner, H., Rolnick, D.: Galileo: Learning global & local features of many remote sensing modalities. In: Proceedings of the 42nd International Conference on Machine Learning (ICML). Proceedings of Machine Learning Research, vol. 267, pp. 60280–60300 (2025)
37. Waldmann, L., Shah, A., Wang, Y., Lehmann, N., Stewart, A.J., Xiong, Z., Zhu, X.X., Bauer, S., Chuang, J.: Panopticon: Advancing any-sensor foundation models for earth observation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops. pp. 2229–2239 (2025)
38. Wang, C., Bai, X., Wang, S., Zhou, J., Ren, P.: Multiscale visual attention networks for object detection in vhr remote sensing images. IEEE Geoscience and Remote Sensing Letters **16**(2), 310–314 (2018)
39. Wang, G., Xie, Y., Jiang, Y., Mandlekar, A., Xiao, C., Zhu, Y., Fan, L., Anandkumar, A.: Voyager: An open-ended embodied agent with large language models. arXiv preprint arXiv:2305.16291 (2023)
40. Wang, Y., Xiong, Z., Liu, C., Stewart, A.J., Dujardin, T., Bountos, N.I., Zavras, A., Gerken, F., Papoutsis, I., Leal-Taixé, L., et al.: Towards a unified copernicus foundation model for earth vision. arXiv preprint arXiv:2503.11849 (2025)
41. Waqas Zamir, S., Arora, A., Gupta, A., Khan, S., Sun, G., Shahbaz Khan, F., Zhu, F., Shao, L., Xia, G.S., Bai, X.: isaid: A large-scale dataset for instance segmentation in aerial images. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops. pp. 28–37 (2019)
42. Wei, Z., Yao, W., Liu, Y., Zhang, W., Lu, Q., Qiu, L., Yu, C., Xu, P., Zhang, C., Yin, B., et al.: Webagent-r1: Training web agents via end-to-end multi-turn reinforcement learning. arXiv preprint arXiv:2505.16421 (2025)

43. Xia, G.S., Bai, X., Ding, J., Zhu, Z., Belongie, S., Luo, J., Datcu, M., Pelillo, M., Zhang, L.: Dota: A large-scale dataset for object detection in aerial images. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3974–3983 (2018)
44. Xia, G.S., Hu, J., Hu, F., Shi, B., Bai, X., Zhong, Y., Zhang, L., Lu, X.: Aid: A benchmark data set for performance evaluation of aerial scene classification. IEEE Transactions on Geoscience and Remote Sensing **55**(7), 3965–3981 (2017)
45. Xu, W., Yu, Z., Mu, B., Wei, Z., Zhang, Y., Li, G., Wang, J., Peng, M.: Rs-agent: Automating remote sensing tasks through intelligent agent. arXiv preprint arXiv:2406.07089 (2024)
46. Yang, A., Li, A., Yang, B., Zhang, B., Hui, B., Zheng, B., Yu, B., Gao, C., Huang, C., Lv, C., et al.: Qwen3 technical report. arXiv preprint arXiv:2505.09388 (2025)
47. Yao, S., Zhao, J., Yu, D., Du, N., Shafran, I., Narasimhan, K., Cao, Y.: React: Synergizing reasoning and acting in language models. In: The Eleventh International Conference on Learning Representations (ICLR) (2023)
48. Zhang, P., Xu, H., Tian, T., Gao, P., Li, L., Zhao, T., Zhang, N., Tian, J.: Sefepnet: Scale expansion and feature enhancement pyramid network for sar aircraft detection with small sample dataset. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing **15**, 3365–3375 (2022)
49. Zhang, T., Zhang, X., Li, J., Xu, X., Wang, B., Zhan, X., Xu, Y., Ke, X., Zeng, T., Su, H., et al.: Sar ship detection dataset (ssdd): Official release and comprehensive data analysis. Remote Sensing **13**(18), 3690 (2021)
50. Zheng, Z., Yang, M., Hong, J., Zhao, C., Xu, G., Yang, L., Shen, C., Yu, X.: Deepeyes: Incentivizing "thinking with images" via reinforcement learning. arXiv preprint arXiv:2505.14362 (2025)
51. Zhu, H., Chen, X., Dai, W., Fu, K., Ye, Q., Jiao, J.: Orientation robust object detection in aerial images using deep convolutional neural network. In: 2015 IEEE international conference on image processing (ICIP). pp. 3735–3739. IEEE (2015)