SENTINELGEO: EVALUATING GEOGRAPHIC GROUNDING IN VISION-LANGUAGE MODELS FROM SATELLITE IMAGERY

Anonymous authors
Paper under double-blind review

ABSTRACT

Vision-language models encode substantial geographic knowledge, yet it remains unclear how reliably they can apply it to a new observation and state a location claim that can be checked. We introduce **SentinelGeo**, a benchmark that turns single-image overhead geolocation into a falsifiable test of geographic grounding. It pairs 11,210 globally distributed, coordinate-labeled Sentinel-2 observations with an image-only protocol in which a model sees one satellite image, with no metadata, retrieval candidates, or tools, and must return a latitude-longitude prediction together with the evidence it relies on. Unlike existing overhead benchmarks, SentinelGeo scores both, auditing each rationale against the image and the true region. Across 18 systems, predictions depart clearly from random but remain coarse, with the best model missing by a median of 1,151 km, while audited rationales score poorly for narrowing the search toward the target region. A calibrated open baseline shows that lightweight task adaptation improves coordinate accuracy, including on a geographically held-out region, without a comparable change in rationale scores. SentinelGeo thus separates what models can place from what they can justify. The dataset and evaluation code will be publicly released.

1 INTRODUCTION

Foundation models trained at internet scale acquire extensive knowledge about the physical world, yet it remains unclear how reliably they can use this knowledge to interpret new visual observations. Spatial structure provides a concrete and measurable setting for studying this question, since geographic position, proximity, and scale constrain what can be inferred from observation. Recent work Gurnee & Tegmark (2024) suggests that Large Language Models (LLMs) form internal representations of geographic space, encoding cities and landmarks across multiple scales. Yet evidence for internal geographic representations does not establish that such representations can be grounded in new visual observations. The key challenge is therefore not merely to ask whether models store geographic facts, or whether geographic variables can be decoded from their representations, but whether latent spatial knowledge can be conditioned on new observations, externalized through coordinate-level predictions, and tested through falsifiable spatial claims.

To turn this question into an evaluation, a model must be asked not only what it knows about the world, but whether it can use visual evidence to locate a new observation within it. Image geolocation provides such a test: it maps visual evidence onto global geographic structure and yields a directly falsifiable spatial prediction. Ground-level geolocation Hays & Efros (2008); Li et al. (2024); Talreja et al. (2026); Vivanco Cepeda et al. (2023); Weyand et al. (2016) often benefits from semantic and cultural cues such as signs, road markings, storefronts, and vehicles. Overhead satellite imagery reduces many of these cues and shifts the burden of localization toward physical and spatial structure. From above, location can be constrained by land cover, terrain, hydrology, agriculture, and urban form. In this sense, single-image overhead geolocation is not merely a remote-sensing application, but a behavioral diagnostic of spatial grounding in general-purpose vision-language systems. Satellite products are distributed with their coordinates, and this is what makes the test possible: every observation carries a precise, globally consistent ground truth, and withholding it turns latent geographic knowledge into a prediction that can be scored geodesically. Geolocation is thus the probe rather than the product. The probed capability still matters where coordinates are missing

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

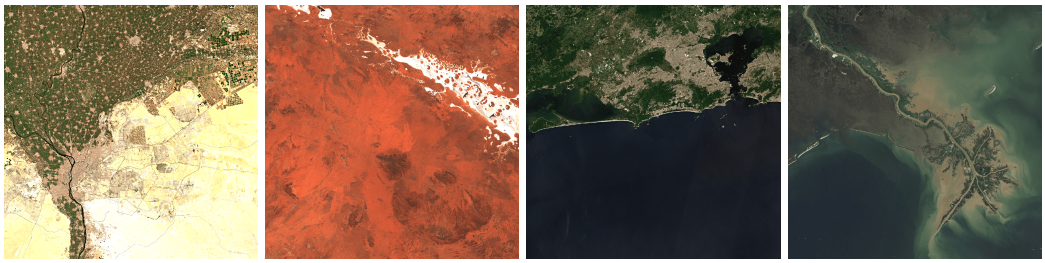


Figure 1: **Overhead imagery as a geolocation test.** Examples from SentinelGeo illustrate both visually distinctive landmarks and spatial cues such as terrain, hydrology, urban form, and land-cover structure. The task requires models to map such evidence to structured geographic coordinate claims without external tools. Inputs are 1024×1024 RGB images covering approximately 110×110 km, with location metadata withheld. These distinctive examples illustrate the task rather than overall benchmark difficulty.

or untrustworthy, as for imagery that circulates without its original metadata, recorded metadata that needs an independent consistency check, historical aerial and satellite archives Piermattei et al. (2026), and astronaut photography Berton et al. (2024). There, a coarse image-only estimate narrows where reference-based registration or retrieval must search Heisig & Simmen (2021), and SentinelGeo measures how well current models provide it.

However, satellite geolocation poses a challenge: visible overhead patterns rarely correspond to unique locations, as similar landscapes, settlement structures, and coastlines recur worldwide. A point prediction's distance error alone cannot establish whether its explanation is supported by the image. This ambiguity is not a flaw of the task; rather, it makes overhead geolocation a useful test of whether models can combine visual constraints with geographic priors, and motivates evaluation beyond final distance error, including whether predictions are regionally consistent and supported by image evidence.

Motivated by this perspective, we formulate single-image satellite geolocation as a behavioral evaluation of geographic grounding: given a single overhead observation, a model must infer a latitude-longitude location, making its spatial claim both globally constrained and directly falsifiable. We instantiate this evaluation through SentinelGeo, a globally distributed benchmark for testing whether multimodal models can map overhead visual evidence to real geographic coordinates. Unlike closed-set remote-sensing classification or region-limited retrieval, SentinelGeo requires prediction in a global coordinate space and evaluates errors with geodesic metrics across countries, continents, and spatial scales. Our contributions are threefold:

- **Coordinate-level evaluation of spatial grounding.** We formulate single-image overhead geolocation as a falsifiable benchmark for testing whether vision-language models can externalize spatial world knowledge as structured latitude-longitude predictions. We curate SentinelGeo, a global geolocation benchmark built from 11,210 coordinate-labeled satellite images.
- **Reasoning-aware coordinate evaluation.** We design an image-only evaluation protocol that measures geodesic coordinate accuracy alongside a post-hoc audit of generated rationales, examining how spatial precision and the visual support of explicit explanations align or diverge across models.
- **Empirical findings on spatial grounding and calibration.** We evaluate a broad set of contemporary vision-language models on SentinelGeo. The experiments show coarse localization and limited visual support in generated rationales. Task adaptation improves localization, including on a region withheld from task-specific training.

2 Related Work

2.1 Geographic Representations and Visual Geolocation

Spatial knowledge in foundation models has been studied both as an internal representation and as an external prediction problem. Representation-level studies show that geographic variables,

such as city and landmark locations, can be decoded from language-model activations Gurnee & Tegmark (2024), suggesting that models may organize some world knowledge spatially. Visual geolocation provides one such external interface. Classical and modern systems localize images through geotagged-image retrieval Hays & Efros (2008), classification over global geocells Weyand et al. (2016), contrastive image-location alignment Vivanco Cepeda et al. (2023), satellite-derived location embeddings Klemmer et al. (2025), and VLM-based geolocation or reasoning pipelines Li et al. (2024); Wang et al. (2024); Yerramilli et al. (2025); Talreja et al. (2026); Li et al. (2025a); Qian et al. (2025). Overhead images have also been geolocalized directly by matching them to georeferenced overhead references Tian et al. (2020), cross-view methods match ground-level queries to aerial or satellite references Workman et al. (2015); Zhu et al. (2021), and EarthLoc Berton et al. (2024) localizes astronaut photographs by retrieval against satellite imagery. These methods require a georeferenced reference database, whereas SentinelGeo asks what a model can infer without one. SentinelGeo builds on this line of work but repurposes geolocation as an evaluation of coordinate-level spatial grounding.

2.2 Earth-Observation Foundation Models and Benchmarks

Earth-observation foundation models have made overhead imagery increasingly accessible to general-purpose multimodal systems. Recent surveys Huang et al. (2025b); Hong et al. (2026); Zhu et al. (2026) trace the development of multimodal systems for satellite and aerial imagery. Systems including SkySense Guo et al. (2024), RSGPT Hu et al. (2025), GeoChat Kuckreja et al. (2024), LHRS-Bot-Nova Li et al. (2025c), and SkyEyeGPT Zhan et al. (2025) support increasingly diverse forms of overhead-image understanding. Benchmarks including GEOBench-VLM Danish et al. (2025), VLEO-Bench Zhang (2025); Zhang & Wang (2024), ChatEarthBench Yuan et al. (2026), OmniEarth Fu et al. (2026), and COREval An et al. (2024) evaluate these capabilities. These benchmarks emphasize semantic and task-specific evaluation: models are judged by whether they recognize scene content, respond correctly to questions or instructions, or reason over annotated image properties. SentinelGeo targets a different output space. It treats the overhead image not only as content to be understood, but as evidence for placing an observation in the global coordinate system.

2.3 Evidence Use and Reasoning Reliability

Evaluation of model reasoning increasingly distinguishes final-answer correctness from the reliability of the rationale that supports it. Chain-of-thought and explanation-based prompting can make model behavior more interpretable, but generated rationales may also be post-hoc, incomplete, or inconsistent with the evidence used to produce the answer Wei et al. (2022); Turpin et al. (2023); Lanham et al. (2023). This concern is especially important for spatial grounding. A coordinate prediction may be close to the target because of a broad geographic prior, while the accompanying explanation cites cues that are generic, weakly discriminative, or absent from the image. Recent geospatial evaluations reflect this broader shift toward process-sensitive evaluation. Earth-observation benchmarks have begun to test whether models reason reliably about geometry, scale, and distribution shift in Earth imagery Xu et al. (2026); Liu et al. (2026); Li et al. (2025b). In visual geolocation, process-oriented studies such as GeoChain Yerramilli et al. (2025) and GeoRC Talreja et al. (2026) further show that plausible location predictions can be supported by brittle or hallucinated reasoning. SentinelGeo extends this line of evaluation to coordinate-level overhead geolocation by treating the predicted location and its supporting evidence as jointly evaluable claims. In addition to measuring geodesic error, it asks whether the model's rationale provides image-grounded evidence that meaningfully constrains the prediction in geographic space. This post-hoc audit concerns explicit explanations, not the model's internal reasoning process.

3 SentinelGeo Evaluation Framework

3.1 Evaluative Scope and Design Rationale

SentinelGeo evaluates overhead geolocation as a behavioral test of image-conditioned spatial grounding. It asks whether a model can use an overhead observation as grounding evidence for a coordinate-

level claim about the Earth, using only the image and a fixed prompt, without location metadata or external tools. This framing leads to three evaluation questions:

1. **Coordinate-level externalization.** How accurately can a model predict latitude and longitude from a single overhead image under this protocol?
2. **Post-hoc explanation audit.** What visual inconsistencies and limitations in geographic support does a fixed judge rubric identify in the explicit explanations?
3. **Effects and regional transfer of post-training.** How does task adaptation change localization and explanation-audit scores, and do its effects transfer to the held-out region?

These questions bound the evaluation scope. SentinelGeo supports comparative behavioral claims under a fixed image-only protocol; it does not reveal how geographic knowledge is represented internally, establish that every image is uniquely identifiable, or prove that proprietary systems reason without memorization. Closed-source models are evaluated as black boxes with unknown pretraining exposure, while the calibrated open baseline enables controlled analysis of task-specific exposure and regional transfer.

3.2 Dataset Construction

SentinelGeo contains 11,210 coordinate-labeled overhead observations constructed from Sentinel-2 Level-2A imagery. Each example uses a 1024×1024 RGB preview from the visual asset of a Sentinel-2 observation; the corresponding native Level-2A product has $10\,\mathrm{m}$ ground sampling distance and a 10980×10980 pixel footprint, covering approximately $109.8\,\mathrm{km} \times 109.8\,\mathrm{km}$ on the ground. The preview represents this tile coverage; $10\,\mathrm{m}$ is the native product resolution, not the preview pixel size. We pair each image with the corresponding tile-center latitude and longitude. Provenance metadata such as acquisition and source identifiers are retained for documentation, splitting, and evaluation, but are never provided to models at inference time.

A central challenge is separating *invalid ambiguity* from *geographic ambiguity*. Some crops are difficult because overhead evidence genuinely underdetermines the location, while others are difficult only because the image is corrupted, occluded, or visually unusable. SentinelGeo removes the latter while preserving the former. Automated filtering excludes severe cloud contamination, missing footprints, large acquisition borders, invalid pixels, and other image-quality failures. We then apply manual inspection to remove only residual cases where interpretable surface evidence is largely absent, such as images dominated by open ocean, heavy cloud or haze, extreme edge artifacts, or visually uninterpretable surfaces. This curation is not used to select easy or iconic locations. Ambiguous but usable croplands, drylands, river systems, peri-urban areas, coastal margins, and rural settlement patterns are retained because such ambiguity is part of the evaluation.

We partition SentinelGeo by evaluative function at the image-record level. A 5,600-sample *calibration split* is used only for the open diagnostic baseline described in Section 3.4. Within this split, 600 samples provide supervised task-interface examples with rationale-and-coordinate targets, while the remaining 5,000 are image-coordinate pairs used for reward-based calibration. A separate 1,000-sample validation split supports training monitoring and model selection, and a 1,000-sample benchmark split is reserved for standardized cross-model evaluation. We additionally define a 421-sample South Asian Subcontinent Holdout, excluded from all task-specific calibration, validation, and model selection and used only for evaluation. The remaining 3,189 examples form an auxiliary reserve split released with the dataset but not used in any reported calibration, model selection, or evaluation. Additional details are provided in Appendix A.

3.3 Benchmark Protocol and Metrics

SentinelGeo operationalizes the image-only setting with a uniform prompt and shared evaluation script. For each sample, the model is prompted to produce a geolocation rationale followed by a structured latitude-longitude prediction in 1-degree coordinate tokens. The final answer must contain two parseable tags, one for latitude and one for longitude, such as `<LAT><N><37></LAT>` and `<LON><W><122></LON>`. The rationale is used for reasoning analysis, while the parsed coordinate tokens are converted into signed latitude-longitude values and treated as the model's falsifiable spatial claim. Dataset metadata remain hidden from the model and are used only by the evaluation code.

Reported outputs follow service-call and parse-failure retries, plus completion scans. Unresolved records remain in the manifest; they are excluded from distance and GWPS averages and counted as administrative non-matches. Full prompts, parsing rules, and parse-failure handling are provided in Appendix B.1.

The primary quantitative measure is geodesic error. For a predicted coordinate $\hat{y}_i = (\hat{\phi}_i, \hat{\lambda}_i)$ and ground-truth coordinate $y_i = (\phi_i, \lambda_i)$, we compute the Haversine distance $d_i = D_{\mathrm{hav}}(\hat{y}_i, y_i)$. The main and holdout tables use $d_{i,\mathrm{fp}} = \max(0, d_i - 77.64\,\mathrm{km})$, subtracting the tile half-diagonal as a footprint approximation, not an exact boundary distance. We report both mean and median footprint-aware distance: the mean captures large global failures, while the median better reflects typical performance under the long-tailed error distribution of worldwide geolocation. We also report city and country matches as auxiliary administrative indicators. These use offline reverse geocoding with custom footprint-aware tolerances: a point-based match, a prediction within 77.64 km of the tile center, or a matching sampled footprint label within 155.28 km is accepted. A 5×5 grid supplies the footprint labels; city labels include country codes (Appendix B.3.1). We treat these indicators as secondary to geodesic error because nearest-place assignment can be unstable near borders, coastlines, and sparsely populated regions.

To complement these distance statistics, we define a benchmark-specific Gaussian distance-utility score, which we call the **Gaussian Weighted Positioning Score (GWPS)**:

$$\mathrm{GWPS} = \frac{1}{|\mathcal{V}|} \sum_{i \in \mathcal{V}} \exp\left(-\frac{1}{2} \left(\frac{d_{i,\mathrm{fp}}}{201.7} \right)^2 \right),$$

where $\mathcal{V}$ contains parseable predictions and $d_{i,\mathrm{fp}}$ is in kilometers, following the valid-prediction averaging convention in Appendix B.1.2. We anchor the decay at 750 km, a country-level threshold used in geolocation tasks Jia et al. (2025); Tang et al. (2026). As a benchmark design choice, we set $\sigma = 201.7$ km so that $u(750) \approx 0.001$.

Beyond coordinate accuracy, SentinelGeo audits whether each rationale is supported by the image and by the ground-truth geographic context. We use GPT-5.4 as a separate rubric-based judge. Given the overhead image, model rationale, predicted coordinate, and ground-truth coordinate, the judge returns a fixed JSON object with four 0–5 integer scores: Image-Reasoning Consistency **(IRC)** for visual faithfulness, Image Inferability / Evidence Sufficiency **(IRI)** for whether the image contains sufficient geographically discriminative evidence, Reasoning-Ground-Truth Consistency **(RGC)** for compatibility with the true region, and Reasoning-to-Geographic-Target Entailment **(RTE-GEO)** for whether the rationale narrows the search toward the target. The judge is used as a structured consistency checker rather than an unconstrained preference model: it compares explicit rationale claims against image evidence and ground-truth metadata under a fixed rubric. Since GPT-family systems are also evaluated in our benchmark, these scores are interpreted only as post-hoc diagnostics of evidence use and geographic plausibility. IRI concerns image evidence sufficiency, not an independent measure of model reasoning ability. Full judge prompts and metric details are provided in Appendix B.1.

3.4 Calibration Baseline

Black-box model comparisons reveal system-level behavior but do not make task-specific adaptation observable. Proprietary models such as GPT-5.2 OpenAI (2025) and Gemini-3.1 Flash Google DeepMind (2026) may differ in pretraining data, instruction tuning, geospatial exposure, and output formatting, so their scores are useful for auditing current systems but difficult to interpret mechanistically. We therefore include an open calibrated baseline built from Qwen2.5-VL-7B-Instruct Bai et al. (2025). Here calibration denotes task adaptation, not probabilistic calibration. This baseline provides a diagnostic test of whether lightweight calibration can stabilize the geolocation interface and improve coordinate-level spatial contraction under the SentinelGeo protocol.

Calibration has two stages. A 600-sample supervised subset aligns the response interface, using teacher-generated rationales conditioned on the image and ground-truth coordinate, paired with verified ground-truth coordinate targets and including manually repaired rationales (Table 6). A larger 5,000-sample image-coordinate subset is then used for reward-based post-training, where geodesic error provides the optimization signal for spatial contraction. The South Asian Subcontinent Holdout is excluded from all supervised calibration, reward-based calibration, validation, and model selection.

Results on this split therefore test whether calibration transfers to a post-training-unseen region rather than directly fitting regional examples. Further details are provided in Appendix B.2.

4 Results and Findings

We evaluate 18 vision-language models (VLMs) on SentinelGeo, spanning proprietary general-purpose models OpenAI (2025; 2026b;a; 2024); Anthropic (2025); Mallick & Kilpatrick (2025); Google (2025); Google DeepMind (2026); xAI (2025), open general-purpose models Zhipu AI (2026); Moonshot AI (2026); Meta AI (2025); Mistral AI (2026); Qwen Team (2026), remote-sensing or geospatial VLMs Li et al. (2026); Wang et al. (2025), and the open calibration pair built on Qwen2.5-VL-7B Bai et al. (2025). The results are organized around the three evaluation questions introduced in Section 3.1: coordinate-level externalization, post-hoc explanation audit, and the effects and regional transfer of post-training.

4.1 Coordinate-Level Externalization

Table 1 addresses whether models can turn overhead observations into metrically meaningful coordinate claims. The random coordinate reference yields a median error of 9590.85 km and only 3.5% country match, whereas most evaluated models achieve substantially lower errors and higher country-level agreement, indicating that their predictions differ systematically from this random reference. However, absolute errors remain large: Gemini 3.1 Flash obtains the best median error, 1151.15 km, but still reaches only 8.8% footprint-aware city match.

Finding 1: Non-random but coarse coordinate externalization. Current VLMs exhibit non-random coordinate-level behavior relative to the random coordinate reference, but this behavior remains coarse-grained: the best systems often identify plausible countries or regions, while city-level and fine-grained coordinate accuracy remain weak.

Table 1: **Footprint-aware coordinate-level performance on SentinelGeo.** City and Country use the footprint-aware matching rule, and distance uses $d_{\mathrm{fp}} = \max(0, d - 77.64\,\mathrm{km})$. Model references are given in the text.

Model	City ↑	Country ↑	Mean (km) ↓	Median (km) ↓	GWPS ↑
Calibration pair					
Qwen2.5-VL-7B	0.2%	13.5%	6064.03	5309.93	0.0194
Qwen2.5-VL-7B (calibrated)	2.3%	27.3%	3781.32	1861.67	0.0898
Other open general-domain VLMs					
GLM-4.6v	2.8%	19.6%	5659.80	4033.49	0.0708
Kimi K2.5	3.7%	23.8%	5467.13	3781.92	0.0886
Llama4 Maverick	2.0%	20.9%	6163.17	5084.16	0.0539
Mistral Large	0.2%	3.8%	8550.62	9036.38	0.0064
Qwen3.5 397B A17B	3.5%	19.8%	6006.18	5223.82	0.0874
Remote-sensing / geospatial VLMs					
GeoReason	0.5%	10.4%	5904.06	5092.55	0.0312
GeoZero	0.5%	10.0%	6031.01	4958.93	0.0192
Proprietary general-domain VLMs					
GPT-5.2	6.5%	27.4%	4592.15	2647.09	0.1447
GPT-5 Mini	5.9%	30.6%	4417.28	2396.74	0.1496
GPT-5.4 Mini	1.4%	18.6%	5293.21	3666.88	0.0544
GPT-4o Mini	2.3%	22.6%	5703.17	3872.89	0.0698
Claude Haiku 4.5	2.1%	16.2%	6266.64	5395.99	0.0420
Gemini 2.0 Flash Lite	5.4%	36.7%	3430.98	1457.36	0.1638
Gemini 2.5 Flash Lite	5.8%	35.6%	3649.36	1449.37	0.1695
Gemini 3.1 Flash	8.8%	44.4%	3154.67	1151.15	0.2236
Grok 4 Fast	2.6%	20.3%	5802.16	4667.59	0.0741
Reference baseline					
Random Coordinate	0.1%	3.5%	9831.46	9590.85	0.0000

We further analyze the full raw geodesic-error distribution using distance bins, with the complete distribution reported in Appendix C. These bins use the raw distance d rather than d_{fp}. The threshold view reinforces Table 1: only 2.5% of random-coordinate predictions fall within 2,000 km of the target, compared with 25.8% for Qwen2.5-VL-7B and 50.5% after calibration. Gemini 3.1 Flash reaches 60.4%, the strongest threshold result among evaluated systems, but nearly 40% of its predictions still miss by more than 2,000 km. Thus, current gains are best interpreted as coarse spatial contraction rather than fine-grained localization. Because predictions use 1-degree coordinate tokens and each tile spans roughly 110 km, the bins cannot speak to placement below the degree grid.

A second pattern is the strong performance of the Gemini family, with Gemini 3.1 Flash and the Flash Lite variants ranking near the top. Since proprietary training mixtures are unobserved, SentinelGeo cannot disentangle stronger spatial reasoning from differential geospatial exposure, so we treat these results as black-box system behavior. This motivates the open calibration pair as a more controlled comparison: starting from the same Qwen2.5-VL-7B base model, task calibration increases country accuracy from 13.5% to 27.3% and reduces median error from 5309.93 km to 1861.67 km. We defer stronger interpretation to Section 4.3, where the South Asian Subcontinent Holdout tests whether this improvement transfers beyond regions seen during task-specific calibration.

4.2 Post-Hoc Explanation Audit and Failure Modes

The second evaluation examines whether a model's stated rationale substantiates its coordinate claim. Table 2 shows that rationale reliability is limited even among models with relatively strong coordinate performance. All models remain well below the top of the 0–5 scale. The weakest dimension is RTE-GEO, where rationales seldom provide evidence that meaningfully narrows the search toward the target region. Because RGC and RTE-GEO compare the rationale against the ground-truth region, a low score on either can reflect an incorrect predicted location rather than an incorrect reading of the image. Rationale reliability therefore serves as a distinct diagnostic rather than a proxy for coordinate accuracy.

Finding 2: Audited rationales are often descriptive rather than localizing. Current models can describe visible overhead patterns, but their rationales frequently lack the geographic specificity needed to justify the emitted coordinate. Plausible scene descriptions therefore do not necessarily constitute reliable evidence-conditioned geolocation.

The calibration pair exposes a sharper separation between coordinate behavior and rationale reliability. Table 1 shows that task calibration substantially improves Qwen2.5-VL-7B's coordinate performance, whereas Table 2 shows little corresponding change: IRI rises from 1.73 to 1.91, while IRC, RGC and RTE-GEO each move by at most 0.04. Given the limited size of the calibration set, this pattern is better interpreted as calibration of the coordinate interface than as learning a new remote-sensing geolocation capability from scratch. Calibration makes location claims easier to externalize through the required coordinate format, but it does not appear to induce a corresponding improvement in evidence-conditioned reasoning.

The low RTE-GEO scores reflect several observed failure modes, whose frequency we did not annotate. First, models produce visually faithful but geographically generic descriptions, such as noting cropland, aridity, or urban density without identifying cues that distinguish one region from many visually similar regions. Second, models sometimes over-index on broad priors, mapping deserts, deltas, or dense cities to familiar regions even when the image lacks region-specific evidence. Third, rationales may hallucinate local evidence, such as named infrastructure, climate zones, or settlement patterns that are not visible in the image. Finally, the emitted coordinate can be inconsistent with the stated rationale, placing a prediction in a different country or biome from the explanation. These errors explain why coordinate proximity and rationale reliability diverge: a model may contract the search space for weak reasons, while a visually plausible rationale may still fail to justify a coordinate-level claim.

4.3 Adaptation Effects and Regional Transfer

The third evaluation question concerns the geographic scope of task calibration. The remaining question is whether this coordinate improvement is confined to regions observed during task-specific training, or whether it transfers to a geographically excluded region. We evaluate this using the

Table 2: **Rationale-reliability scores on SentinelGeo.** Scores are diagnostic 0–5 measures of image consistency (IRC), image evidence sufficiency (IRI), ground-truth compatibility (RGC), and target-region narrowing (RTE-GEO). IRI scores the evidence available in the image rather than the model's reasoning, so the four dimensions are reported separately.

Model	IRC ↑	IRI ↑	RGC ↑	RTE-GEO ↑
Open calibration pair				
Qwen2.5-VL-7B	1.67	1.73	1.47	0.74
Qwen2.5-VL-7B (calibrated)	1.62	1.91	1.43	0.71
Other open general-domain VLMs				
GLM-4.6v	2.15	1.88	1.56	0.83
Kimi K2.5	1.98	1.97	1.17	0.74
Llama4 Maverick	2.33	1.97	1.92	0.89
Mistral Large	0.75	1.04	0.34	0.11
Qwen3.5 397B A17B	1.77	1.86	0.94	0.58
Remote-sensing / geospatial VLMs				
GeoReason	3.05	2.35	1.64	0.88
GeoZero	2.75	2.24	1.34	0.65
Proprietary general-domain VLMs				
GPT-5.2	2.50	2.19	1.60	1.01
GPT-5 Mini	2.52	2.19	1.69	1.09
GPT-5.4 Mini	2.61	2.10	1.61	0.89
GPT-4o Mini	2.28	1.95	1.82	1.02
Claude Haiku 4.5	1.98	1.94	1.14	0.60
Gemini 2.0 Flash Lite	2.26	1.90	**2.13**	1.44
Gemini 2.5 Flash Lite	2.19	2.01	2.04	1.32
Gemini 3.1 Flash	**3.17**	**2.43**	1.69	**1.63**
Grok 4 Fast	2.12	1.98	1.33	0.66

421-image South Asian Subcontinent Holdout, which is excluded from supervised calibration, reward-based calibration, validation, and model selection.

Table 3: **South Asian subcontinent holdout results.** Coordinate performance uses the footprint-aware metric and matching rule on the 421-image geographically held-out region excluded from task-specific calibration, validation, and model selection. Proprietary models serve as black-box reference points; the controlled comparison is between the open base model and its task-calibrated variant.

Coordinate-level performance					
Model	City ↑	Country ↑	Mean (km) ↓	Median (km) ↓	GWPS ↑
Qwen2.5-VL-7B	0.2%	3.3%	6229.16	5132.61	0.0148
Qwen2.5-VL-7B (calibrated)	0.7%	8.1%	3821.61	3685.59	0.0347
Gemini 3.1 Flash	9.0%	54.9%	2118.84	653.28	0.2582
GPT-5.4 Mini	0.7%	20.2%	5030.74	4020.37	0.0572
Claude Haiku 4.5	1.0%	7.8%	7356.44	7776.43	0.0342
Rationale-reliability diagnostics					
Model		IRC ↑	IRI ↑	RGC ↑	RTE-GEO ↑
Qwen2.5-VL-7B		1.65	1.64	0.86	0.30
Qwen2.5-VL-7B (calibrated)		1.58	1.95	1.13	0.30
Gemini 3.1 Flash		3.03	2.35	1.97	1.76
GPT-5.4 Mini		2.55	2.14	1.50	0.85
Claude Haiku 4.5		2.09	2.00	1.05	0.52

Table 3 reports both coordinate performance and rationale-reliability diagnostics on the held-out region. The controlled comparison is the calibration pair, where the calibrated model is initialized from Qwen2.5-VL-7B and differs only by the task-specific calibration applied in our setup. On the holdout, task calibration increases country accuracy from 3.3% to 8.1%, reduces mean error

from 6229.16 km to 3821.61 km, and reduces median error from 5132.61 km to 3685.59 km. These changes show that the effect of calibration remains visible on the held-out region, but they do not yet establish fine-grained localization or reliable explanatory grounding. Whether this region appeared in the base model's pretraining is unobserved, so the comparison does not rule out prior exposure.

Finding 3: Calibration effects persist on the held-out region, but only coarsely. On the held-out region, the calibrated model still reduces mean and median error and raises country agreement relative to its base model. Current models can externalize some spatial structure under overhead evidence, but they do not yet provide fine-grained localization or reliable evidence-conditioned spatial reasoning.

The limits of this held-out transfer are as informative as the improvement itself. The calibrated model still has 0.7% footprint-aware city agreement, a median error of 3685.59 km, and an unchanged RTE-GEO score of 0.30. Calibration changes where the model predicts more than it changes the quality of the rationale supporting that prediction, reinforcing the separation between coordinate-level externalization and evidence-conditioned reasoning. Gemini 3.1 Flash performs strongly on the holdout, but its training data, geospatial exposure, and alignment procedures are unobserved. These rows show how current proprietary systems behave under SentinelGeo, not why they behave that way. The image-only setting also serves as a control for tool-augmented systems: adding a retrieval pipeline or Google Search to Gemini 3.1 Flash lowered its country agreement from 44.4% to about 38% and did not reduce its median error (Appendix C.4). Supplementary controls for grid formulation, tool use, prompt and sampling robustness, Sentinel-2 product level, and a ground-level cross-domain setting are reported in Appendix C, with implementation details in Appendix B.

Limitations. SentinelGeo is an evaluation benchmark rather than an operational geolocation system. A single overhead image often underdetermines global location, so the benchmark measures how models combine ambiguous visual constraints with geographic priors rather than whether coordinates are visually determined. Predictions use a 1-degree coordinate grid, which bounds attainable precision; regional transfer is tested on a single held-out region; proprietary models have unknown geospatial exposure; and dataset construction may reflect geographic, seasonal, cloud-filtering, and land-cover biases. Each prediction uses one image at one acquisition time without tools; multi-temporal, multi-scale, and tool-augmented settings supply evidence beyond a single observation and are complementary evaluations rather than part of this benchmark. The model-judged rationale-reliability audits are diagnostic rather than definitive evidence of causal reasoning, and are not human-validated.

5 Conclusion

We introduced SentinelGeo, a benchmark that turns single-image overhead geolocation into a falsifiable test of geographic grounding in vision-language models, and that, unlike existing overhead benchmarks, scores the coordinate prediction together with the evidence a model states for it. Current systems place predictions far from random but remain coarse, with the best model reaching only 8.8% footprint-aware city agreement, and their audited rationales score poorly for narrowing the search toward the target region. Lightweight task calibration of an open baseline improves the coordinate metrics, including on a geographically held-out region, without a comparable change in the rationale scores, so coordinate accuracy and explanatory support move largely independently. These remain behavioral measurements under one protocol, and finer coordinate interfaces, human-validated rationale audits, and held-out designs spanning more regions would sharpen what such a protocol can establish. We release SentinelGeo to support that work.

Ethics Statement

SentinelGeo is built from public medium-resolution Sentinel-2 Earth-observation imagery and contains no human-subject or street-level data. Coordinate prediction from overhead imagery is nonetheless dual-use, and could be misused for surveillance, sensitive-site inference, or targeting. SentinelGeo is intended for model auditing and research evaluation, not for safety-critical or operational deployment, and the released metadata documents provenance, intended use, non-intended use, known biases, and misuse risks.

Reproducibility Statement

The authors will release SentinelGeo as a full dataset archive, a smaller inspection sample, and the accompanying evaluation code. The full archive contains images, coordinate metadata, functional split files, and README documentation for the calibration, validation, benchmark, auxiliary, and held-out partitions, while the smaller sample supports quick inspection without downloading the full archive. The evaluation scripts, prompt templates, coordinate parser, metric implementations, invalid-output handling rules, and the rationale-reliability rubric needed to reproduce the reported evaluations are documented in Appendix B. The dataset release is accompanied by metadata with Responsible AI fields documenting provenance, intended use, non-intended use, known biases, limitations, and misuse risks.

AI Use Statement

The authors used generative AI tools in three ways. First, GPT-5.4 serves as an LLM-as-a-judge for the rationale-reliability scores (Section 3.3 and Appendix B.1.3). To limit judge bias, it scores four heads independently under a fixed rubric that instructs it to be strict and evidence-based, and each output must match a fixed JSON schema with bounded 0–5 scores. These scores do not affect any coordinate prediction or the primary metrics, which are computed deterministically from the parsed coordinates. They are treated only as a supplementary post-hoc diagnostic, because the judge sees the ground-truth coordinate, belongs to the same model family as some evaluated systems, and has not been validated against human annotation. Second, a teacher model generated the rationales in the 600-sample supervised calibration set, which were then manually screened and repaired, while coordinate supervision always comes from verified ground truth (Appendix B.2). Third, AI assistants were used for language polishing and editing of the manuscript. The authors reviewed all AI-assisted text and take full responsibility for the content of this paper.

References

Xiao An, Jiaxing Sun, Zihan Gui, and Wei He. COREval: A comprehensive and objective benchmark for evaluating the remote sensing capabilities of large vision-language models. *arXiv preprint arXiv:2411.18145*, 2024.

Anthropic. Introducing Claude Haiku 4.5. https://www.anthropic.com/news/claude-haiku-4-5, 2025. Accessed: 2026-04-27.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025.

Gabriele Berton, Alex Stoken, Barbara Caputo, and Carlo Masone. EarthLoc: Astronaut photography localization by indexing earth from space. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, pp. 12754–12764, 2024.

Muhammad Sohail Danish, Muhammad Akhtar Munir, Syed Roshaan Ali Shah, Kartik Kuckreja, Fahad Shahbaz Khan, Paolo Fraccaro, Alexandre Lacoste, and Salman Khan. GEOBench-VLM: Benchmarking vision-language models for geospatial tasks. In *Proc. IEEE Int. Conf. Comp. Vis.*, pp. 7132–7142, 2025. doi: 10.1109/ICCV51701.2025.00670.

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

Getty Research Institute. Getty thesaurus of geographic names online. https://www.getty.edu/research/tools/vocabularies/tgn/, 2017.

Google. We're expanding our Gemini 2.5 family of models. https://blog.google/products-and-platforms/products/gemini/gemini-2-5-model-family-expands/, 2025. Accessed: 2026-04-27.

Google DeepMind. Gemini 3.1 Flash. https://deepmind.google/models/gemini/flash/, 2026. Accessed: 2026-04-27.

Xin Guo, Jiangwei Lao, Bo Dang, Yingying Zhang, Lei Yu, Lixiang Ru, Liheng Zhong, Ziyuan Huang, Kang Wu, Dingxiang Hu, Huimei He, Jian Wang, Jingdong Chen, Ming Yang, Yongjun Zhang, and Yansheng Li. SkySense: A multi-modal remote sensing foundation model towards universal interpretation for earth observation imagery. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, pp. 27672–27683, 2024. doi: 10.1109/CVPR52733.2024.02613.

Wes Gurnee and Max Tegmark. Language models represent space and time. In *Proc. Int. Conf. Learn. Representations*, 2024.

James Hays and Alexei A Efros. Im2gps: estimating geographic information from a single image. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, pp. 1–8. IEEE, 2008.

Holger Heisig and Jean-Luc Simmen. Re-engineering the past: Countrywide geo-referencing of archival aerial imagery. *PFG – Journal of Photogrammetry, Remote Sensing and Geoinformation Science*, 89:487–503, 2021. doi: 10.1007/s41064-021-00162-z.

Danfeng Hong, Chenyu Li, Xuyang Li, Gustau Camps-Valls, and Jocelyn Chanussot. Foundation models in remote sensing: Evolving from unimodality to multimodality. *IEEE Geoscience and Remote Sensing Magazine*, 14(1):1–20, 2026. doi: 10.1109/MGRS.2026.3669086.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Liang Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. In *Proc. Int. Conf. Learn. Representations*, 2022.

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

Fei Huang, Fan Wu, Zeqing Zhang, Qihao Wang, Long Zhang, Grant Michael Boquet, and Hongyang Chen. GeoGPT-RAG technical report. *arXiv preprint arXiv:2509.09686*, 2025a.

Ziyue Huang, Hongxi Yan, Qiqi Zhan, Shuai Yang, Mingming Zhang, Chenkai Zhang, Yiming Lei, Zeming Liu, Qingjie Liu, and Yunhong Wang. A survey on remote sensing foundation models: From vision to multimodality. *arXiv preprint arXiv:2503.22081*, 2025b.

Pengyue Jia, Seongheon Park, Song Gao, Xiangyu Zhao, and Sharon Li. GeoRanker: Distance-aware ranking for worldwide image geolocalization. In *Proc. Advances in Neural Inf. Process. Syst.*, volume 38, pp. 20100–20126, 2025. doi: 10.52202/085713-0596.

Konstantin Klemmer, Esther Rolf, Caleb Robinson, Lester Mackey, and Marc Rußwurm. SatCLIP: Global, general-purpose location embeddings with satellite imagery. *Proc. AAAI Conf. Artificial Intell.*, 39(4): 4347–4355, 2025. doi: 10.1609/aaai.v39i4.32457.

Kartik Kuckreja, Muhammad Sohail Danish, Muzammal Naseer, Abhijit Das, Salman Khan, and Fahad Shahbaz Khan. GeoChat: Grounded large vision-language model for remote sensing. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, pp. 27831–27840, 2024. doi: 10.1109/CVPR52733.2024.02629.

Tamera Lanham, Anna Chen, Ansh Radhakrishnan, Benoit Steiner, Carson Denison, Danny Hernandez, Dustin Li, Esin Durmus, Evan Hubinger, Jackson Kernion, et al. Measuring faithfulness in chain-of-thought reasoning. *arXiv preprint arXiv:2307.13702*, 2023.

Ling Li, Yu Ye, Bingchuan Jiang, and Wei Zeng. GeoReasoner: Geo-localization with reasoning in street views using a large vision-language model. In *Proc. Int. Conf. Mach. Learn.*, volume 235 of *Proceedings of Machine Learning Research*, pp. 29222–29233. PMLR, 2024.

Lingyao Li, Runlong Yu, Qikai Hu, Bowei Li, Min Deng, Yang Zhou, and Xiaowei Jia. From pixels to places: A systematic benchmark for evaluating image geolocalization ability in large language models. *arXiv preprint arXiv:2508.01608*, 2025a.

Wenshuai Li, Xiantai Xiang, Zixiao Wen, Guangyao Zhou, Ben Niu, Feng Wang, Lijia Huang, Qiantong Wang, and Yuxin Hu. GeoReason: Aligning thinking and answering in remote sensing vision-language models via logical consistency reinforcement learning. *arXiv preprint arXiv:2601.04118*, 2026. doi: 10.48550/arXiv.2601.04118.

Xiang Li, Yong Tao, Siyuan Zhang, Siwei Liu, Zhitong Xiong, Chunbo Luo, Lu Liu, Mykola Pechenizkiy, Xiao Xiang Zhu, and Tianjin Huang. REOBench: Benchmarking robustness of earth observation foundation models. In *Proc. Advances in Neural Inf. Process. Syst.*, volume 38, pp. 68313–68334, 2025b. doi: 10.52202/085713-2052.

Zhenshi Li, Dilxat Muhtar, Feng Gu, Yanglangxing He, Xueliang Zhang, Pengfeng Xiao, Guangjun He, and Xiaoxiang Zhu. LHRS-Bot-Nova: Improved multimodal large language model for remote sensing vision-language interpretation. *ISPRS Journal of Photogrammetry and Remote Sensing*, 227:539–550, 2025c. doi: 10.1016/j.isprsjprs.2025.06.003.

Yixin Liu, Yifan Zhang, Hao Wang, Zhen Chen, Jun Li, Xinyi Wang, et al. VLRS-Bench: A vision-language reasoning benchmark for remote sensing. *arXiv preprint arXiv:2602.07045*, 2026.

Shrestha Basu Mallick and Logan Kilpatrick. Gemini 2.0: Flash, flash-lite and pro. https://developers.googleblog.com/en/gemini-2-family-expands/, 2025. Accessed: 2026-04-27.

Meta AI. The Llama 4 herd: The beginning of a new era of natively multimodal intelligence. https://ai.meta.com/blog/llama-4-multimodal-intelligence/, 2025. Accessed: 2026-04-27.

Mistral AI. Models – from cloud to edge. https://mistral.ai/models, 2026. Accessed: 2026-04-27.

Moonshot AI. Kimi K2.5: Open visual agentic model for real work. https://www.kimi.com/ai-models/kimi-k2-5, 2026. Accessed: 2026-04-27.

OpenAI. GPT-4o mini: Advancing cost-efficient intelligence. https://openai.com/index/gpt-4o-mini-advancing-cost-efficient-intelligence/, 2024. Accessed: 2026-04-27.

OpenAI. Introducing GPT-5.2. https://openai.com/index/introducing-gpt-5-2/, 2025. Accessed: 2026-04-27.

OpenAI. Introducing GPT-5.4 mini and nano. https://openai.com/index/introducing-gpt-5-4-mini-and-nano/, 2026a. Accessed: 2026-04-27.

OpenAI. GPT-5 mini model. https://developers.openai.com/api/docs/models/gpt-5-mini, 2026b. Accessed: 2026-04-27.

Livia Piermattei, Robert McNabb, Melanie Elias, Camillo Ressl, Amaury Dehecq, Luc Girod, Thomas Dewez, and Anette Eltner. Unlocking the past: A review of digital processing of historical aerial and satellite stereo analog imagery for geoscience applications. *IEEE Geoscience and Remote Sensing Magazine*, 2026. doi: 10.1109/MGRS.2025.3645144.

Zhaofang Qian, Hardy Chen, Zeyu Wang, Li Zhang, Zijun Wang, Xiaoke Huang, Hui Liu, Xianfeng Tang, Zeyu Zheng, Haoqin Tu, Cihang Xie, and Yuyin Zhou. Where on earth? a vision-language benchmark for probing model geolocation skills across scales. *arXiv preprint arXiv:2510.10880*, 2025.

Qwen Team. Qwen3.5: Towards native multimodal agents. https://qwen.ai/blog?id=qwen3.5, 2026. Accessed: 2026-04-27.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. DeepSeekMath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

Mohit Talreja, Joshua Diao, Jim Thannikary James, Radu Casapu, Tejas Santanam, Ethan Mendes, Alan Ritter, Wei Xu, and James Hays. GeoRC: A benchmark for geolocation reasoning chains. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 40540–40564. Association for Computational Linguistics, 2026. doi: 10.18653/v1/2026.acl-long.1883.

Yong Tang, Jianhua Gong, Yi Li, Jieping Zhou, and Jun Sun. Synergizing retrieval and CoT reasoning via spatial consensus for worldwide visual geo-localization. *ISPRS International Journal of Geo-Information*, 15(4):163, 2026. doi: 10.3390/ijgi15040163.

Yuxin Tian, Xueqing Deng, Yi Zhu, and Shawn Newsam. Cross-time and orientation-invariant overhead image geolocalization using deep local features. In *Proc. Winter Conf. Appl. of Comp. Vis.*, pp. 2501–2510, 2020.

Miles Turpin, Julian Michael, Ethan Perez, and Samuel Bowman. Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting. In *Proc. Advances in Neural Inf. Process. Syst.*, volume 36, pp. 74952–74965, 2023.

Vicente Vivanco Cepeda, Gaurav Kumar Nayak, and Mubarak Shah. Geoclip: Clip-inspired alignment between locations and images for effective worldwide geo-localization. In *Proc. Advances in Neural Inf. Process. Syst.*, volume 36, pp. 8690–8701, 2023.

Leandro von Werra, Younes Belkada, Lewis Tunstall, Edward Beeching, Tristan Thrush, Nathan Lambert, Shengyi Huang, Kashif Rasul, and Quentin Gallouédec. TRL: Transformer reinforcement learning. https://github.com/huggingface/trl, 2020.

Di Wang, Shunyu Liu, Wentao Jiang, Fengxiang Wang, Yi Liu, Xiaolei Qin, Zhiming Luo, Chaoyang Zhou, Haonan Guo, Jing Zhang, Bo Du, Dacheng Tao, and Liangpei Zhang. GeoZero: Incentivizing reasoning from scratch on geospatial scenes. *arXiv preprint arXiv:2511.22645*, 2025. doi: 10.48550/arXiv.2511.22645.

Zhiqiang Wang, Dejia Xu, Rana Muhammad Shahroz Khan, Yanbin Lin, Zhiwen Fan, and Xingquan Zhu. LLMGeo: Benchmarking large language models on image geolocation in-the-wild. *arXiv preprint arXiv:2405.20363*, 2024.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. In *Proc. Advances in Neural Inf. Process. Syst.*, volume 35, pp. 24824–24837, 2022.

Tobias Weyand, Ilya Kostrikov, and James Philbin. PlaNet: Photo geolocation with convolutional neural networks. In *Proc. Eur. Conf. Comp. Vis.*, pp. 37–55. Springer, 2016.

Scott Workman, Richard Souvenir, and Nathan Jacobs. Wide-area image geolocalization with aerial reference imagery. In *Proc. IEEE Int. Conf. Comp. Vis.*, pp. 3961–3969, 2015. doi: 10.1109/ICCV.2015.451.

xAI. Grok 4 Fast. `https://x.ai/news/grok-4-fast`, 2025. Accessed: 2026-04-27.

Zelin Xu, Yupu Zhang, Saugat Adhikari, Saiful Islam, Tingsong Xiao, Zibo Liu, Shigang Chen, Da Yan, and Zhe Jiang. Earthspatialbench: Benchmarking spatial reasoning capabilities of multimodal llms on earth imagery. *arXiv preprint arXiv:2602.15918*, 2026.

Sahiti Yerramilli, Nilay Pande, Rynaa Grover, and Jayant Sravan Tamarapalli. GeoChain: Multimodal chain-of-thought for geographic reasoning. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 23624–23639. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.findings-emnlp.1284.

Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Weinan Dai, Tiantian Fan, Gaohong Liu, Lingjun Liu, et al. DAPO: An open-source LLM reinforcement learning system at scale. *arXiv preprint arXiv:2503.14476*, 2025.

Zhenghang Yuan, Zhitong Xiong, Thomas Dujardin, Xiang Li, Lichao Mou, and Xiao Xiang Zhu. ChatEarth-Bench: Benchmarking multimodal large language models for earth observation. *IEEE Geoscience and Remote Sensing Magazine*, 14(2):64–84, 2026. doi: 10.1109/MGRS.2026.3650840.

Yang Zhan, Zhitong Xiong, and Yuan Yuan. Skyeyegpt: Unifying remote sensing vision-language tasks via instruction tuning with large language model. *ISPRS Journal of Photogrammetry and Remote Sensing*, 221:64–77, 2025. doi: 10.1016/j.isprsjprs.2025.01.020.

Chenhui Zhang. VLEO-Bench: A framework to evaluate vision-language models for earth observation applications. Master's thesis, Massachusetts Institute of Technology, 2025.

Chenhui Zhang and Sherrie Wang. Good at captioning, bad at counting: Benchmarking GPT-4V on earth observation data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pp. 7839–7849, 2024. doi: 10.1109/CVPRW63382.2024.00780.

Zhipu AI. Model overview. `https://docs.bigmodel.cn/cn/guide/start/model-overview`, 2026. Accessed: 2026-04-27.

Sijie Zhu, Taojiannan Yang, and Chen Chen. VIGOR: Cross-view image geo-localization beyond one-to-one retrieval. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, pp. 3640–3649, 2021. doi: 10.1109/CVPR46437.2021.00364.

Xiao Xiang Zhu, Zhitong Xiong, Yi Wang, Adam J. Stewart, Konrad Heidler, Yuanyuan Wang, Zhenghang Yuan, Thomas Dujardin, Qingsong Xu, and Yilei Shi. On the foundations of earth foundation models. *Commun. Earth Environ.*, 7(1):103, 2026. doi: 10.1038/s43247-025-03127-x.

APPENDIX

Due to space limitations in the main paper, we defer additional technical details, reproducibility material, and extended analyses to this appendix. The purpose of the appendix is not to introduce separate claims beyond the main paper, but to document the assumptions, implementation choices, and supporting evidence behind SentinelGeo as a coordinate-level evaluation framework for geographic grounding in vision-language models. We first describe the dataset construction pipeline, split

definitions, South Asian Subcontinent Holdout, and released artifacts. We then consolidate all experimental settings and implementation details in a dedicated Implementation section, covering the main benchmark, rationale-reliability audit, calibrated baseline, and supplementary controls. We further report supplementary quantitative results, distributional analyses, and qualitative failure cases that complement the main findings. The appendix is organized as follows.

Appendix Contents

A Dataset Construction **14**

A.1 Data Source and Sample Construction 14

A.2 Filtering and Quality Control 16

A.3 Functional Splits and Released Files 16

B Implementation **18**

B.1 Main Benchmark Evaluation Protocol and Metrics 18

B.2 Calibrated Qwen2.5-VL Baseline Details 23

B.3 Supplementary Experiment Implementations 26

C Additional Results **28**

C.1 Additional Quantitative Results 28

C.2 Footprint-Aware Evaluation and Alternative Grid Formulations 29

C.3 Stage-by-Stage Calibration Ablation 29

C.4 External Tools and the Agentic Setting 30

C.5 Prompt and Sampling Robustness 30

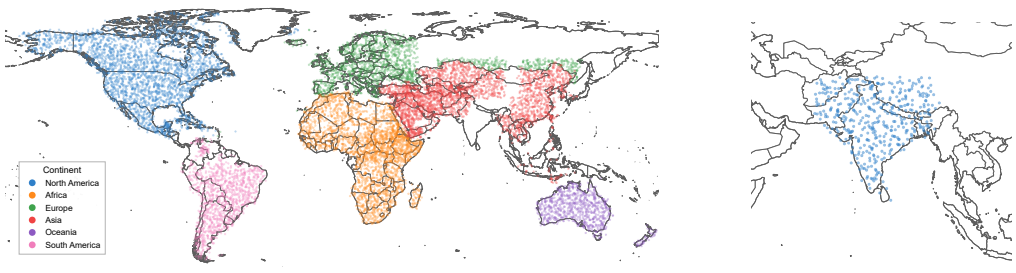
C.6 Level-1C (TOA) versus Level-2A (BOA) 31

C.7 Cross-Domain GeoGuessr Control 32

C.8 Qualitative Examples 32

A Dataset Construction

A.1 Data Source and Sample Construction



(a) SentinelGeo Distribution (b) Holdout Distribution

Figure 2: **Spatial coverage of SentinelGeo.** Left: all samples outside the holdout. Right: South Asian Subcontinent Holdout. The holdout is excluded from all task-specific calibration, validation, and model selection and is used only for regional evaluation.

SentinelGeo is constructed from public Sentinel-2 Level-2A overhead imagery accessed through the Microsoft Planetary Computer STAC API. We use `sentinel-2-l2a` as the sole imagery collection and use the RGB preview of the corresponding `visual` asset. Each example consists of an overhead RGB image, a ground-truth latitude-longitude coordinate, and acquisition metadata used for documentation and evaluation. During model evaluation, only the image and benchmark prompt are provided to the model; coordinates, timestamps, tile identifiers, cloud metadata, region labels, and other geographic metadata are hidden from the model. To obtain a geographically diverse pool, we query the STAC catalog over a set of continental and regional bounding boxes rather than a single contiguous study area. For each query region, we restrict the search to recent observations from the last 18 months, sort candidate items from most recent to oldest acquisition time, and retrieve up to 500 candidate entries per query before downstream quality control. This procedure is intended to cover a broad range of overhead visual evidence, including terrain, land cover, hydrology, agricultural structure, coastlines, settlement morphology, and climate-related spatial patterns, while avoiding dependence on any single region or acquisition date.

We additionally audit acquisition time to bound when the benchmark observations came into existence. The earliest acquisition is 2024-10-14 15:12:01 UTC and the latest is 2025-10-25 05:29:21 UTC, spanning 375 days (12.3 months). Table 4 reports the yearly and monthly acquisition distribution.

Table 4: **Yearly and monthly acquisition distribution.** The audit summarizes when the curated SentinelGeo observations were acquired. Shares are rounded to one decimal place and may not sum exactly to 100%.

Year	Share	Year–month	Share
2024	7.6%	2024-10	1.9%
2025	92.4%	2024-11	3.0%
		2024-12	2.7%
		2025-01	4.5%
		2025-02	4.7%
		2025-03	15.9%
		2025-04	43.7%
		2025-05	0.9%
		2025-06	1.5%
		2025-07	4.5%
		2025-08	4.9%
		2025-09	7.1%
		2025-10	4.9%

Acquisitions are concentrated in March and April 2025, which together account for 59.6% of the observations, so the benchmark is not temporally uniform. This acquisition-time record bounds when the benchmark observations came into existence, but it does not establish private pretraining membership for proprietary systems, which cannot be verified directly.

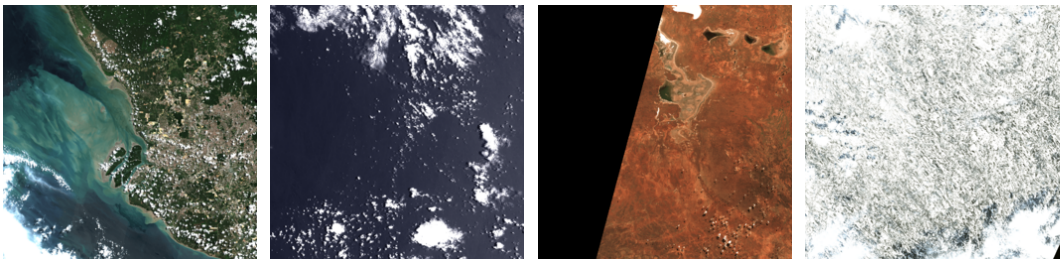
For each retained STAC item, we download the RGB preview from the `visual` asset and store it as a 1024×1024 image. We associate the image with the tile-center coordinate used as the ground-truth location for geodesic evaluation. The resulting sample format supports a strict image-conditioned geolocation protocol: the image provides the visible overhead evidence, while the coordinate is used only after inference to compute error metrics and auxiliary geographic matches. This separation between model-visible input and evaluation-only metadata is central to SentinelGeo's role as a coordinate-level evaluation framework rather than a metadata-retrieval benchmark. Figure 2 summarizes the spatial coverage of the curated dataset and the South Asian Subcontinent Holdout. The full pool covers a broad range of geographic and land-cover conditions, while the holdout region is reserved for evaluation and excluded from all task-specific calibration, validation, and model selection.

As shown in Figure 2b, the South Asian Subcontinent Holdout is designed to assess whether task-specific calibration transfers beyond regions observed during post-training. It contains the 421 samples whose tile centers fall within $6.8°$–$35.7°$N and $64.4°$–$92.4°$E, a region covering the South Asian subcontinent and parts of neighboring countries, including India, Pakistan, Bangladesh, Nepal, Bhutan, Sri Lanka, Afghanistan, and western China. This split is excluded from SFT calibration, reward-based calibration, validation, and model selection. Therefore, the holdout controls task-

specific exposure, testing whether calibration improves the coordinate metrics in a region not seen during SentinelGeo post-training. It does not rule out geographic knowledge acquired during pretraining, especially for closed-source systems whose training mixtures are unknown.

A.2 Filtering and Quality Control

The raw STAC query results contain candidate scenes that are not always suitable for image-conditioned geolocation evaluation. We therefore apply a multi-stage quality-control pipeline before constructing the final SentinelGeo pool. At the source level, we require `eo:cloud_cover < 1` (cloud cover below 1%) to reduce large-scale cloud occlusion. After downloading the RGB previews, we further remove incomplete or corrupted samples, including images with large black-border artifacts caused by missing tile coverage or partial scene overlap.



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

Figure 3: **Representative filtered-out samples.** From left to right: cloud contamination, ocean-dominated scene, black-border artifact, and snow or haze obscuring the surface.

We also perform manual screening to remove observations whose visible surface content is not usable for the intended evaluation. Excluded cases include scenes severely affected by residual cloud cover, haze or snow that obscures the surface, large acquisition artifacts, or scenes dominated by open ocean or otherwise weakly interpretable surface content. Representative filtered-out examples are shown in Figure 3. These examples illustrate the types of unusable observations removed during quality control, rather than examples that are merely difficult to localize.

Importantly, the filtering procedure is not intended to select only visually distinctive or easy geolocation cases. Many retained images remain ambiguous, low-cue, or compatible with multiple possible world regions. Such ambiguity is central to SentinelGeo's evaluative design: the benchmark tests how models combine imperfect overhead visual evidence with internal geographic priors to produce coordinate-level spatial claims. Accordingly, the manual screening step is designed to remove unusable observations, not difficult observations. Visually ambiguous but interpretable scenes are retained whenever they contain meaningful surface evidence such as terrain structure, coastline geometry, hydrological patterns, agricultural texture, settlement form, or land-cover composition.

After quality control, the curated pool is partitioned into functional splits for calibration, validation, benchmark evaluation, and held-out regional analysis. The split construction and released metadata are described in the next subsection. Figure 4 shows representative retained SentinelGeo samples, illustrating the diversity of overhead cues and the range of visual ambiguity preserved after quality control.

A.3 Functional Splits and Released Files

After quality control, the curated SentinelGeo pool contains 11,210 RGB overhead images. We partition this pool according to functional evaluation roles rather than treating all examples as a single homogeneous split. Table 5 summarizes the resulting partitions. The SFT and reward subsets of the calibration split are used only for constructing the Qwen2.5-VL-7B (calibrated) diagnostic baseline; the validation split is used for training monitoring and model selection; the SentinelGeo benchmark split is used for the main cross-model evaluation; and the South Asian Subcontinent Holdout is reserved for held-out regional analysis. The auxiliary reserve is released for inspection and future analysis but is not used in the main reported results.



Figure 4: **Representative retained SentinelGeo samples.** The examples illustrate the diversity of overhead visual evidence preserved after quality control, including terrain structure, coastlines, hydrology, agricultural texture, settlement morphology, desert and arid landscapes, snow-covered regions, and low-cue but interpretable scenes. These samples are not intended to be uniquely identifiable from pixels alone; rather, they show the range of ambiguous grounding signals used to evaluate image-conditioned coordinate prediction.

Table 5: **Functional splits in SentinelGeo.** Splits are defined by their role in calibration, validation, and evaluation. The South Asian Subcontinent Holdout is excluded from all task-specific calibration, validation, and model selection.

Split	Size	Role
Calibration split (SFT subset)	600	SFT for the calibrated baseline
Calibration split (reward subset)	5,000	reward-based coordinate calibration
Validation split	1,000	training monitoring and model selection
Benchmark split	1,000	main cross-model evaluation
South Asian Subcontinent Holdout	421	held-out regional evaluation
Auxiliary reserve	3,189	released reserve examples, not used in reported results
Total	11,210	

Split construction and spatial unit. The split program uses fixed random seed 42 and samples without replacement at the *image-record* level. It allocates 600 records to SFT, 5,000 to distance-reward calibration, 1,000 to validation, and then 1,000 records from the remaining pool to the benchmark split; the released manifests record the final functional assignment for every sample. Table 5 therefore reports image counts. The South Asian Subcontinent Holdout is defined separately by geographic region and is excluded from task-specific SFT, reward calibration, validation, and checkpoint selection. Because the holdout and the other five partitions are disjoint, the seed-42 allocation draws from the 10,789 non-holdout records, and the remaining 3,189 records form the auxiliary reserve. Records are assigned individually rather than by location, so geographically adjacent observations can fall into different non-holdout splits, and only the holdout is separated by region.

Table 6: **SFT screening disposition for the final 600-example set.** Every example enters the accepted set through a documented screening path.

Acceptance path	Examples	Share
Accepted on the first GPT-5.2 generation	408	68.0%
Accepted after an automated GPT-5.2 retry	90	15.0%
Accepted after documented manual repair and re-audit	102	17.0%
Final accepted set	600	100.0%

Table 6 lists how each example entered the final set. Coordinate supervision always comes from verified ground-truth tokens rather than teacher-predicted coordinates, separating teacher assistance in rationale construction from the source of location supervision. The teacher did see the ground-truth coordinate when writing each rationale, however, so the rationales remain conditioned on the target location even after the teacher's own coordinates are removed.

B IMPLEMENTATION

This section consolidates the implementation and reproducibility details for the main benchmark, rationale-reliability audit, calibrated baseline, and all supplementary experiments. The result-focused appendix sections report outcomes and interpretation only.

B.1 MAIN BENCHMARK EVALUATION PROTOCOL AND METRICS

B.1.1 PROMPTING TEMPLATE

SentinelGeo uses a unified image-conditioned evaluation protocol for both API-based models and local models. For each benchmark example, the model receives a single overhead RGB image together with a fixed geolocation instruction. The evaluation is conducted without auxiliary geographic side information: models are not given metadata, retrieval results, candidate regions, or access to external geolocation tools. The benchmark therefore evaluates whether a model can use the overhead image and its internal geographic knowledge to produce a coordinate-level spatial claim.

Unlike standard free-form geolocation prompts, SentinelGeo requires the final coordinate to be emitted in a structured discrete-token format. This design improves parseability across heterogeneous model APIs and local VLM checkpoints, and it makes coordinate extraction deterministic. The model is instructed to write a rationale before the final answer, but the final prediction must end with two latitude and longitude token lines. These tokens encode integer-degree latitude and longitude values.

Inference prompt. The following system and user prompts are used for benchmark inference and are reproduced verbatim. Minor API-specific message formatting differs across backends, but the instruction content and required final coordinate format are fixed.

> **System prompt.**
>
> You are a geolocation reasoning expert. Infer the location strictly from visible image evidence. Output the final location using discrete latitude and longitude tokens only.
>
> At the very end, output exactly two lines:
>
> `<LAT><D_lat><V_lat></LAT>`
> `<LON><D_lon><V_lon></LON>`
>
> `D_lat` is `N` or `S`; `V_lat` is a 2-digit magnitude for 1-degree bins, e.g., 37° is written as `<37>`. `D_lon` is `E` or `W`; `V_lon` is a 3-digit magnitude for 1-degree bins, e.g., 123° is written as `<123>`.
>
> Examples:
>
> `<LAT><N><37></LAT>`
> `<LON><W><122></LON>`
>
> `<LAT><S><34></LAT>`
> `<LON><E><151></LON>`
>
> `<LAT><N><13></LAT>`
> `<LON><E><100></LON>`
>
> You need write reasoning before the final two lines.
>
> **User prompt.**
>
> You are participating in a geolocation challenge based on satellite images. Carefully analyze the image and reason about which country/region it is in and why. Instead of outputting continuous latitude and longitude, you must use DISCRETE tokens at the end.
>
> Now analyze the given satellite image. Write your reasoning first. Then end your answer with exactly:
>
> `<LAT>...</LAT>`
> `<LON>...</LON>`

Main-inference configuration. For the formal benchmark runs, the shared generation configuration uses temperature 0.8 and a maximum output length of 4,096 tokens. API image inputs are converted to RGB, resized only when necessary so that the longest side does not exceed 1,024 pixels, and JPEG encoded with quality 85 before transmission. The main benchmark output interface is the structured integer-degree latitude/longitude token contract above; continuous decimal coordinates are used only in the dedicated output-interface robustness control (Appendix B.3.5) and are parsed by the corresponding decimal-coordinate parser. The functional splits are drawn with seed 42 (Appendix A.3); each supplementary subset in Appendix B.3.1–B.3.7 states its own seed.

Retry and failure handling. We distinguish service-call failures from coordinate-parse failures. A failed service request is retried up to 50 times. If a response is returned but does not yield a valid latitude–longitude pair, the sample is regenerated up to 50 times, using temperature 0.8 for regeneration. Samples still unresolved after the attempt limit remain in the evaluation manifest and are marked as invalid predictions. After the main pass, the runner checks for missing records and unresolved samples and performs up to three completion scans. This procedure is shared by API and local backends, and for unresolved samples the raw output is still recorded but no distance is computed.

Coordinate-token format. The final latitude token has the form

$$\texttt{<LAT><D_lat><V_lat></LAT>},$$

where $\texttt{D_lat} \in \{\texttt{N}, \texttt{S}\}$ denotes the hemisphere and $\texttt{V_lat}$ is an integer magnitude in degrees. The final longitude token has the form

$$\texttt{<LON><D_lon><V_lon></LON>},$$

where $\texttt{D_lon} \in \{\texttt{E}, \texttt{W}\}$ denotes the hemisphere and $\texttt{V_lon}$ is an integer magnitude in degrees. The parser converts these tokens into signed coordinates by assigning positive values to N and E, and negative values to S and W. Because the magnitudes are integers, the parsed prediction is an integer-degree coordinate rather than a continuous decimal-degree coordinate.

Deterministic parsing. We use deterministic regular-expression parsers for latitude and longitude. The evaluator searches the raw model output for a <LAT>...</LAT> block and a <LON>...</LON> block. Within the latitude block, it searches for an uppercase direction token <N> or <S> and a one- or two-digit magnitude token. Within the longitude block, it searches for an uppercase direction token <E> or <W> and a two- or three-digit magnitude token. If either the direction token or the magnitude token is missing, the corresponding coordinate is treated as unparseable. The parser is thus slightly more permissive than the prompt, which requests exactly two latitude digits and three longitude digits: it also accepts one-digit latitudes and two-digit longitudes, and it does not require the direction token to precede the magnitude token within a block.

After extraction, the latitude magnitude is clipped to the valid range $[0, 90]$, and the longitude magnitude is clipped to $[0, 180]$. The signed coordinate is then computed from the extracted hemisphere direction. A prediction is considered parseable only when both latitude and longitude are successfully extracted. Since the prompt instructs the model to put the final answer in the last two lines, the parsed tag pair is interpreted as the final coordinate prediction.

Backend consistency. The same system prompt, user prompt, coordinate-token format, and parsing logic are used for API models and local models. API models are queried through an OpenAI-compatible chat-completions interface. Local models are evaluated through the corresponding VLM processor and generation interface, with generation stopped after the longitude closing tag when supported. In both cases, the evaluator stores the raw response, the parsed coordinate if available, a parse-success indicator, and the downstream geodesic error. This shared interface ensures that benchmark scores are computed from the same structured coordinate representation across model families.

B.1.2 Quantitative Metrics and Geographic Matching

All quantitative metrics are computed from the parsed latitude-longitude token pair described in Section B.1.1. Because predictions are emitted as integer-degree coordinate tokens, the benchmark

should be interpreted as coordinate-level global localization at integer-degree precision rather than sub-degree geolocation. This discretization improves parseability and enables consistent evaluation across API-based systems and local VLM checkpoints, while still requiring models to make metrically testable spatial claims.

Parse success. For each model response, the evaluator first attempts to parse a latitude token and a longitude token. A prediction is counted as parseable only when both coordinates are successfully extracted. We report the parse success rate as

$$\mathrm{ParseSuccess} = \frac{|\mathcal{V}|}{N},$$

where N is the number of evaluated examples and $\mathcal{V}$ is the set of examples with parseable coordinate predictions. Parse failures are retained in the result file together with the raw model output. They are not assigned artificial coordinates for distance computation. For administrative matching metrics, parse failures count as non-matches.

Result cleaning and evaluation coverage. Before aggregate metrics are computed, saved outputs are de-duplicated by image file name; when duplicate records exist, the last saved record for that image is retained. The finalized main benchmark contains 1,000 unique image records per evaluated model, and the regional holdout contains 421 records per evaluated model. In the saved summaries used for the reported distance statistics, these final retained records have valid coordinate distances, so $|\mathcal{V}| = N$ for the reported main and holdout results. Comparisons are performed on matched manifests: models in the same evaluation use the same image-ID set, not merely the same number of examples. The definitions below are the base point-based rules. Tables 1 and 3 apply them to the footprint-aware distance $d_{i,\mathrm{fp}} = \max(0, d_i - 77.64\,\mathrm{km})$ and use the footprint-aware administrative rule of Appendix B.3.1, and each supplementary table in Appendix C states the convention it uses.

Geodesic distance. For each parseable prediction $i \in \mathcal{V}$, we compute the Haversine distance between the predicted coordinate $(\hat{\phi}_i, \hat{\lambda}_i)$ and the ground-truth coordinate (ϕ_i, λ_i):

$$d_i = 2R \arcsin \sqrt{\sin^2\left(\frac{\phi_i - \hat{\phi}_i}{2}\right) + \cos(\phi_i)\cos(\hat{\phi}_i)\sin^2\left(\frac{\lambda_i - \hat{\lambda}_i}{2}\right)},$$

where latitude and longitude are expressed in radians and R denotes the Earth's radius. Distances are reported in kilometers. We report both mean distance,

$$\mathrm{MeanDist} = \frac{1}{|\mathcal{V}|}\sum_{i\in\mathcal{V}} d_i,$$

and median distance,

$$\mathrm{MedianDist} = \mathrm{median}_{i\in\mathcal{V}}(d_i).$$

Mean distance is sensitive to long-tail failures, while median distance reflects the typical localization error among parseable predictions.

Distance-bin statistics. To expose the full error distribution rather than only aggregate statistics, we also compute the number and percentage of parseable predictions falling into fixed distance bins:

$$[0, 2000], \quad (2000, 4000], \quad \ldots, \quad (16000, 18000], \quad (18000, \infty)$$

kilometers. The complete distance-bin distribution for the main benchmark is reported using the same bins in Appendix C.

Administrative geographic matching. We compute city- and country-level matches using offline reverse geocoding. Both ground-truth and predicted coordinates are mapped to nearest-place records using the same reverse-geocoding database. Country match is computed by exact agreement of the reverse-geocoded country code. City match is computed by normalized agreement of the reverse-geocoded locality name when both coordinates resolve to a non-empty locality. Formally,

$$\mathrm{CountryMatch} = \frac{1}{N}\sum_{i=1}^{N} \mathbb{1}[\mathrm{cc}(\hat{y}_i) = \mathrm{cc}(y_i)],$$

where $\mathrm{cc}(\cdot)$ denotes the reverse-geocoded country code. City match is defined analogously using the reverse-geocoded locality name. These metrics are normalized by the total number of evaluated examples N, so parse failures are counted as non-matches.

Administrative matches are used as auxiliary interpretability metrics rather than as the primary benchmark objective. Reverse geocoding maps coordinates to nearest-place entries and can be ambiguous near borders, coastlines, rural regions, and sparse settlements. We therefore interpret country match as a coarse administrative indicator and city match as an approximate locality-scale indicator, while relying on geodesic distance and GWPS for metric localization quality.

Gaussian Weighted Positioning Score. In addition to distance statistics, we report the Gaussian Weighted Positioning Score (GWPS), a bounded localization-utility metric that assigns smoothly decaying credit as geodesic error increases. For each parseable prediction, the distance utility is

$$u(d_i) = \exp\left[-\frac{1}{2}\left(\frac{d_i}{\sigma}\right)^2\right],$$

and the model-level score is

$$\mathrm{GWPS} = \frac{1}{|\mathcal{V}|}\sum_{i\in\mathcal{V}} u(d_i).$$

We use $\sigma = 201.7$ km, chosen so that $u(750) \approx 0.001$. This choice gives near-zero utility to errors at approximately country-to-subcontinental scale while retaining sensitivity to predictions that approach the target at sub-regional scales. GWPS lies in $[0, 1]$, with larger values indicating better coordinate-level localization. Unlike country or city match, GWPS is independent of administrative boundaries and varies continuously with geodesic error.

Together, these metrics separate output-format reliability, metric localization accuracy, administrative agreement, and smooth distance-based utility. In the main paper, mean and median distance are the primary quantitative indicators, GWPS provides a complementary distance-based utility, city and country match are auxiliary administrative indicators, and distance-bin statistics are reported in the appendix.

B.1.3 Rationale-Reliability Audit Rubric

In addition to coordinate-level metrics, SentinelGeo evaluates whether a model's stated rationale is supported by visible overhead evidence and whether that rationale is geographically useful. This rationale-reliability audit is performed as a post-hoc diagnostic: it does not affect the model's prediction, and it is not used to select or correct coordinates. Instead, it provides a structured way to separate final coordinate accuracy from the quality of the evidence and rationale used to justify the prediction.

For each parseable model response, a separate judge model, GPT-5.4, is given the overhead image, the model's rationale, the parsed predicted coordinate, and the ground-truth coordinate. The judge is instructed to evaluate satellite-image geo-reasoning with a strict rubric and to return only valid JSON. Because the judge sees the ground-truth coordinate, the resulting scores should be interpreted as retrospective diagnostic scores rather than deployment-time confidence estimates.

Judge prompt. The judge uses the following system instruction and user-template structure, reproduced verbatim. The prompt's label "Reasoning-to-GT Entailment" refers to the RTE-GEO head, which this paper expands as Reasoning-to-Geographic-Target Entailment.

> **System prompt.**
>
> You are an expert evaluator for satellite remote sensing geo-reasoning. You evaluate reasoning traces for satellite image geolocation tasks using a strict rubric. You must score four heads independently:
>
> 1. Image-Reasoning Consistency (IRC)
> 2. Image Inferability / Evidence Sufficiency (IRI)
> 3. Reasoning-Ground Truth Consistency (RGC)
> 4. Reasoning-to-GT Entailment (RTE-GEO)
>
> You must be strict, evidence-based, and resistant to fluent nonsense. You must return ONLY valid JSON.

User prompt template.

Evaluate the following satellite geo-reasoning sample.

```
[Satellite Image]
<image>
[Reasoning Trace]
{reasoning}
[Predicted Coordinates]
{pred_coord}
[Ground Truth Coordinates]
{gt_coord}
```

Rubric:

1) IRC (0–5): Is the reasoning faithful to visible image evidence?
2) IRI (0–5): Does the image contain sufficient discriminative evidence?
3) RGC (0–5): Is the reasoning geo-semantically consistent with GT?
4) RTE-GEO (0–5): Does the reasoning help narrow the location toward GT?

Return ONLY JSON.

Scoring heads. The judge assigns four scores, each on a 0–5 scale. The four heads are designed to separate different failure modes: hallucinating image evidence, relying on non-discriminative cues, giving an explanation that conflicts with the true region, and producing a rationale that does not narrow the geographic search.

Table 7: **Rationale-reliability scoring heads.** Scores are assigned independently on a 0–5 scale.

Head	Measures	High score indicates
IRC	Image–Reasoning Consistency	Stated visual cues are visible in the image
IRI	Image Inferability / Evidence Sufficiency	Image cues are geographically discriminative
RGC	Reasoning–Ground-Truth Consistency	Rationale is compatible with the true region
RTE-GEO	Reasoning-to-Geographic-Target Entailment	Rationale helps narrow the search toward GT

IRC measures whether the model's stated evidence is actually supported by the image. A low IRC score indicates hallucinated or unsupported visual claims, such as describing coastlines, rivers, urban patterns, or terrain features that are not visible. A high IRC score indicates that the rationale accurately refers to image-visible evidence.

IRI measures whether the image itself contains sufficient geographically discriminative evidence. A scene dominated by generic vegetation, roads, or rural land can be described faithfully and still remain compatible with many regions. High IRI scores indicate that the visible cues provide meaningful geographic constraints. Although IRI asks about the image, the judge assigns it while reading each model's rationale, so IRI can differ across models on the same images (Table 2). We therefore do not interpret between-model differences in IRI as differences in reasoning quality.

RGC measures whether the rationale is semantically compatible with the ground-truth location. This head asks whether the geographic interpretation of the evidence fits the true region, climate, terrain, land-cover pattern, or settlement context. It can be low even when the rationale is fluent and visually plausible if the inferred geographic story contradicts the true region.

RTE-GEO measures whether the rationale narrows the search toward the ground-truth location. This head is stricter than general plausibility: it evaluates whether the rationale, if followed, would move the prediction toward the target rather than merely producing a generic or globally common description.

JSON schema and parsing. The judge is required to return a JSON object with one score and one justification for each head:

```
{
```

```
  "head_scores": {
    "IRC": {"score": 0, "justification": ""},
    "IRI": {"score": 0, "justification": ""},
    "RGC": {"score": 0, "justification": ""},
    "RTE_GEO": {"score": 0, "justification": ""}
  }
}
```

The evaluator extracts the first JSON object from the judge response and validates that all four heads are present. A judge output is accepted only if each head contains an integer score in the range $[0, 5]$. If the JSON is missing, malformed, incomplete, or contains out-of-range scores, the judge call is retried according to the benchmark configuration. If all judge retries fail, the sample is marked as judge-unparseable and excluded from mean judge-score computation.

Interpretation and limitations. The rationale-reliability audit is intended to complement, not replace, coordinate-level metrics. It is useful for identifying cases where a model predicts a relatively nearby coordinate for weak or hallucinated reasons, as well as cases where a model gives a visually faithful rationale that still fails to narrow the location. Because the scores are produced by an LLM judge and the judge is provided with the ground-truth coordinate, they should not be interpreted as independent human annotations or as proof of causal reasoning, and they have not been validated against human annotation. We use them as structured diagnostic signals for comparing evidence use and geographic plausibility under a fixed rubric. GPT-5.4 also belongs to the same model family as several evaluated systems, which is a further reason to treat these scores only as a supplementary post-hoc diagnostic.

B.2 Calibrated Qwen2.5-VL Baseline Details

B.2.1 Role of the Calibrated Open Baseline

The calibrated Qwen2.5-VL baseline is included as a diagnostic open-model reference rather than as a separate model contribution. Starting from Qwen2.5-VL-7B-Instruct, we apply two lightweight calibration stages: supervised calibration on generated rationale-and-coordinate responses, followed by distance-reward calibration with DAPO. The purpose of this baseline is to test whether the structured coordinate-token interface and image-conditioned geolocation behavior can be calibrated under a controlled post-training setup, and to provide a more interpretable comparison point alongside black-box API models. The South Asian Subcontinent Holdout is excluded from SFT data generation, supervised calibration, DAPO calibration, validation, and model selection; therefore, held-out-region results test transfer beyond task-specific exposure, not absence of pretraining exposure. The two stages are designed to play different roles. SFT is intended to establish the task interface, rationale format, and structured coordinate output from teacher-generated targets that are manually screened and paired with verified coordinates, and the DAPO stage is intended to further reduce geodesic error through the distance reward.

B.2.2 Coordinate-Token Interface and SFT Data Generation

The calibrated Qwen2.5-VL baseline uses the same structured coordinate-token interface as the benchmark evaluation protocol. Latitude and longitude are represented with integer-degree precision. A latitude target is written as

$$\texttt{<LAT><D_lat><V_lat></LAT>},$$

where $\texttt{D_lat} \in \{\texttt{N}, \texttt{S}\}$ denotes the hemisphere and $\texttt{V_lat}$ is a magnitude in the range $0, \ldots, 90$. A longitude target is written as

$$\texttt{<LON><D_lon><V_lon></LON>},$$

where $\texttt{D_lon} \in \{\texttt{E}, \texttt{W}\}$ denotes the hemisphere and $\texttt{V_lon}$ is a magnitude in the range $0, \ldots, 180$. Coordinates are converted to tokens by rounding the absolute latitude or longitude to the nearest integer degree, clipping the magnitude to the valid coordinate range, and encoding the sign through the hemisphere token. For example, a coordinate near $(37.4^\circ, -123.8^\circ)$ is encoded as <LAT><N><37></LAT> and <LON><W><124></LON>.

Table 8: **SFT calibration hyperparameters.** The SFT stage adapts Qwen2.5-VL-7B-Instruct with LoRA using generated rationale-and-coordinate responses. The model was trained on an NVIDIA RTX PRO 6000.

Setting	Value
Base model	Qwen2.5-VL-7B-Instruct
Training samples	600
Epochs	3
Batch size	1
Gradient accumulation	8
Learning rate	2×10^{-5}
Weight decay	0.01
Warmup ratio	0.1
Learning-rate scheduler	cosine
Maximum sequence length	4096
LoRA rank	16
LoRA alpha	32
LoRA dropout	0.05
LoRA target modules	q/k/v/o projections; gate/up/down projections
Precision	bf16

To support stable generation and parsing, we define a compact set of latitude-longitude special tokens. The token set includes the base tags <LAT>, </LAT>, <LON>, </LON>, the hemisphere tags <N>, <S>, <E>, and <W>, together with numeric magnitude tokens for latitude and longitude. The implementation supports one-digit, two-digit, and three-digit numeric formats so that both training targets and model outputs can be parsed robustly. In total, the interface uses 299 custom tokens. Rather than expanding the model vocabulary, we apply a token-ID reuse mechanism that maps these custom tokens onto existing tokenizer IDs and saves the resulting mapping for reuse during later calibration and evaluation. This design keeps the model architecture fixed while making the coordinate-output format explicit and deterministic.

We generate supervised calibration targets before SFT using GPT-5.2 as the teacher model. For each selected example, the teacher receives the overhead image and the ground-truth coordinate and produces an image-conditioned rationale followed by coordinate tokens. Teacher assistance is used only to construct supervised calibration targets; during evaluation, neither the calibrated baseline nor any other model receives ground-truth coordinates, region labels, acquisition metadata, or other geographic side information. Detailed screening and target-construction rules are given in Section B.2.3.

B.2.3 Supervised Calibration with LoRA

The first calibration stage fine-tunes Qwen2.5-VL-7B-Instruct with LoRA on the generated SFT targets described in Section B.2.2. The training data are constructed by matching image records from the SFT subset with the generated response records using the shared image identifier. Each training example consists of a system instruction, a user message containing the overhead image and task prompt (shorter than the benchmark prompt in Appendix B.1.1), and an assistant response containing an image-conditioned rationale followed by the verified latitude and longitude tokens. Rows with empty or malformed generated responses are skipped before training.

During supervised calibration, the model is trained only on the assistant response. The system and user portions of the conversation are used as conditioning context but are masked out of the language-modeling loss. Concretely, the training script first tokenizes the full conversation, then separately tokenizes the system-user prefix to identify the boundary at which the assistant response begins. All prompt tokens and padding tokens are assigned label value -100, so the loss is computed only on the generated rationale and final coordinate tokens. We use parameter-efficient LoRA adaptation Hu et al. (2022) rather than full-model fine-tuning. LoRA modules are applied to the attention projections and MLP projection layers of Qwen2.5-VL-7B-Instruct. The custom latitude-longitude token mapping described in Section B.2.2 is applied to the tokenizer before training, and the resulting token map is

Table 9: **DAPO calibration hyperparameters.** DAPO starts from the SFT LoRA adapter and optimizes the structured coordinate-output behavior with a distance-based reward. The model was trained on an NVIDIA RTX PRO 6000.

Setting	Value
Batch size	1
Gradient accumulation	8
Optimizer steps	625
Learning rate	1×10^{-6}
Weight decay	0.1
Learning-rate scheduler	cosine
Optimizer	AdamW (`adamw_torch`)
Generations per prompt	8
Iterations per generation batch (`num_iterations`)	1
KL coefficient β	0.001
Maximum completion length	1024
Sampling temperature	1.0
Top-p	0.9
Stop string	`</LON>`
Reward fine threshold τ_f	1000 km
Reward truncation distance $d_{\max}$	12742 km
Validation samples	1000
Precision	bf16

saved with the adapter so that later calibration and evaluation use the same coordinate-token interface. Table 8 summarizes the SFT configuration.

SFT target construction. SFT targets are generated with GPT-5.2 for the 600 records of the SFT subset (Appendix A.3), at most three automated retries, and a 4,096-token completion limit. Teacher-emitted coordinates are removed before supervision, and verified ground-truth coordinate tokens are appended only after the rationale passes screening. Screening checks six aspects: (1) every stated visual cue must be observable or conservatively inferable from the image; (2) geographic hypotheses must follow from the stated terrain, hydrology, land cover, agriculture, or settlement evidence; (3) ambiguous scenes must use calibrated uncertainty rather than unsupported landmark certainty; (4) rationales must not reveal or copy the supplied ground-truth coordinate; (5) the cleaned response must satisfy the latitude/longitude output contract; and (6) a second full-set audit checks consistency between the cleaned rationale and target tokens. Responses failing these criteria are regenerated or manually repaired and then re-audited before entering the final set. Gradient checkpointing is enabled, and the saved LoRA configuration contains 47,589,376 trainable parameters, approximately 0.57% of the loaded model parameters.

B.2.4 Distance-Reward Calibration with DAPO

The second calibration stage continues from the SFT LoRA adapter and optimizes the model with a distance-based reward using the GRPO trainer Shao et al. (2024) in TRL von Werra et al. (2020) with its DAPO loss option Yu et al. (2025). We refer to this stage as DAPO after that loss option; it is not intended as a reproduction of every component of the original DAPO recipe. This stage is designed to improve coordinate-level spatial contraction and output parseability under the same structured latitude-longitude token interface used in benchmark evaluation. For each training example, the model receives the overhead image and geolocation prompt, generates a rationale followed by coordinate tokens, and the generated tokens are parsed into a signed latitude-longitude prediction. If either coordinate cannot be parsed, the completion receives zero reward. Otherwise, we compute the Haversine distance d between the parsed prediction and the ground-truth coordinate and assign a piecewise linear reward:

$$r(d) = \begin{cases} 0.5 + 0.5\left(1 - \frac{d}{\tau_f}\right), & d \le \tau_f, \\ 0.5\left(1 - \frac{d-\tau_f}{d_{\max}-\tau_f}\right), & \tau_f < d \le d_{\max}, \\ 0, & d > d_{\max}, \end{cases}$$

where $\tau_f = 1000$ km is the fine-distance threshold and $d_{\max} = 12742$ km is the reward truncation distance (Earth's diameter), beyond which a prediction receives zero reward. The reward is therefore highest for near-target predictions, decreases linearly to 0.5 at 1000 km, decays to zero at $d_{\max}$, and gives no credit to parse failures or extremely distant predictions. This reward is used only during DAPO calibration; all reported benchmark scores are computed separately using the evaluation metrics in Appendix B.1.

DAPO training uses the 5,000-record reward subset of the calibration split and periodically evaluates on the validation split for monitoring. We sample multiple completions per prompt, use the distance reward above as the reward function, and stop generation after the longitude closing tag when possible. Table 9 summarizes the DAPO configuration. The South Asian Subcontinent Holdout is not used during SFT data generation, supervised calibration, DAPO reward training, validation, or checkpoint selection.

DAPO implementation. The reward-calibration stage is implemented with `trl.GRPOTrainer` and `GRPOConfig(loss_type="dapo")`. Each prompt produces eight candidate completions and the trainer performs one optimization iteration per sampled batch (`num_iterations=1`); the KL coefficient is $\beta = 0.001$. A completion that cannot be parsed into a valid coordinate pair, or for which the distance computation is invalid, receives reward 0. With gradient accumulation of 8, the 625 optimizer updates in Table 9 correspond to 5,000 micro-steps.

B.3 Supplementary Experiment Implementations

The following subsections collect the settings and execution details for the supplementary controls reported later in Appendix C.

B.3.1 Footprint-Aware Matching

Because each Sentinel-2 input represents a $109.8\,\mathrm{km} \times 109.8\,\mathrm{km}$ native tile rather than a point observation, we report a footprint-aware distance $d_{\mathrm{fp}} = \max(0, d - 77.64\,\mathrm{km})$, where 77.64 km is the tile half-diagonal. For the administrative indicators, the implementation approximates the geographic footprint around the tile center using a half-side length of 54.9 km and a 5×5 sampling grid. Latitude offsets use 111.32 km per degree; longitude offsets are divided by the cosine of the tile-center latitude, with the cosine lower-bounded at 0.02 for numerical stability. Every sampling point is passed through the same offline `reverse_geocoder` pipeline used for the coordinate-level administrative metrics, producing sampled city and country sets for the tile. An administrative match is retained if the original point-based match is already true; it is also accepted when the predicted point lies within 77.64 km of the tile center, or when the prediction lies within twice the half-diagonal (155.28 km) and its administrative label belongs to the sampled footprint set. City-set matches use a normalized locality name jointly with country code to avoid cross-country collisions between identically named places. This is an approximate footprint-aware evaluation based on center-distance tolerances and sampled place-name coverage; it is not an exact point-in-polygon test on the original Sentinel-2 tile geometry and does not compute the exact shortest distance to the physical tile boundary. Tables 1 and 3 use this convention.

B.3.2 Direct MGRS 100-km Classification

Direct-grid inference details. The direct-grid candidate space is generated independently of model outputs by scanning the global coordinate domain at 0.25-degree increments, converting sampled positions to MGRS 100-km identifiers, and de-duplicating the identifiers to produce a fixed 65,230-class candidate list. Local and API grid-classification runs use the same candidate list and the same benchmark-split manifest. The grid run uses the same image preprocessing as the main benchmark, temperature 0.8, and a 4,096-token output ceiling. The parser accepts only a valid candidate-grid identifier; invalid responses enter the same retry/failure accounting rather than being mapped to a nearest valid class. After retries, every response yields a valid grid identifier. Distance statistics are computed only for predictions for which a valid grid-center coordinate is available.

B.3.3 Stage-by-Stage Calibration Ablation

For the formal stage-by-stage ablation, Base, +SFT, and +DAPO are evaluated on the identical image-ID manifest with the same long-form benchmark prompt, the same coordinate parser, the same retry policy, and the same generation budget. In particular, the SFT checkpoint is evaluated in the benchmark/DAPO prompt mode rather than the shorter SFT-training prompt; otherwise the stage difference would be confounded by prompt wording. Holding these interfaces fixed is a precondition for interpreting the rows as a calibration-stage ablation rather than as three unrelated runs on the same data.

B.3.4 External Tools and Agentic Setting

Tool-use implementation. The GeoGPT-RAG condition is a fixed retrieval pipeline: it rewrites the extracted visual evidence into six geographic retrieval queries, retrieves place-name records from the Getty Thesaurus of Geographic Names Getty Research Institute (2017), and embeds the retrieved content with the GeoEmbedding model released with GeoGPT-RAG Huang et al. (2025a) before final coordinate prediction. The Google-Search condition uses the corresponding Gemini native search-grounding path and requires a search action rather than allowing the model to bypass the tool. Query failures are logged explicitly and do not remove the image from evaluation; the prediction path still terminates with a coordinate attempt. In the reported 1,000-image run, both tool conditions record 1,000/1,000 successful required-query executions. Tool-call failure therefore does not remove predictions from either condition. The tool experiment changes the complete agentic pipeline, not only the presence of a retrieval flag, so its comparison is interpreted as a system-level control rather than a pure one-variable ablation of retrieval.

B.3.5 Prompt Robustness and Repeated Sampling

Subset, generation, and aggregation. The robustness subset is a fixed 200-image sample drawn from the benchmark split with seed 2024 and stored as a shared subset manifest; every model and every condition reads the same image IDs. Condition A is generated three independently sampled times per image, whereas B, C, and D are generated once. These robustness runs use temperature 0.5 and a maximum generation length of 4,096 tokens. For the aggregate Condition-A row, the three predictions for each image are ordered by their ground-truth geodesic distance and the *entire prediction record* associated with the median distance is selected. City, Country, mean distance, median distance, and GWPS for Condition A are then computed from these 200 selected records. Because this selection uses the ground truth, the Condition-A aggregate is an analysis device rather than a deployable decoding rule. Because of the lower temperature, the robustness runs are not directly comparable with Table 1. By contrast, Table 15 treats the three A runs separately; its run-level statistics are not recomputed from the median-record aggregation used in Table 14.

B.3.6 L1C/L2A Matched Rendering

Matched rendering pipeline. Both product levels are rendered from the B04/B03/B02 bands into RGB at 1,024-pixel output resolution using bilinear resampling. Reflectance is converted with quantification value 10,000, the product radiometric offset is applied, and a fixed display stretch of $[0, 0.3]$ with gamma 1.0 is used for both L1C and L2A. Same-scene pairing uses a ± 5 minute acquisition-time window in addition to the tile identity; tile-only fallback is disabled so that a different overpass is not substituted when a temporal match is absent. Rendering equivalence is checked on five paired samples using 0–255 RGB values and an MAE acceptance threshold of 3. Acquisition matching and renderer validation are performed as separate checks.

B.3.7 GeoGuessr-Style Ground-Level Control

Street-view scoring and run manifest. The ground-level control uses a fixed 892-image GeoBench/GeoGuessr-style Street View manifest. Country accuracy is normalized by all 892 samples, while distance and task score are computed for non-refused predictions. The score is

$$s(d) = \begin{cases} 5000, & d_{\mathrm{m}} \le 25, \\ \mathrm{round}\big(5000 \times 0.99866017^{d_{\mathrm{m}}/\mathrm{scale}}\big), & d_{\mathrm{m}} > 25, \end{cases}$$

Table 10: **Distance-bin distribution on the SentinelGeo benchmark.** Bins use the raw Haversine distance d_i rather than $d_{i,\mathrm{fp}}$. Each cell reports the percentage of parseable predictions whose geodesic error falls within the corresponding distance range. Distance bins are reported in kilometers. The table complements the aggregate country-match, mean/median distance, and GWPS results in the main paper by showing whether errors concentrate at regional scales or remain dominated by long-distance failures.

Model	0–2k	2–4k	4–6k	6–8k	8–10k	10–12k	12–14k	14–16k	16–18k	>18k
Calibration pair										
Qwen2.5-VL-7B	25.8%	16.9%	10.1%	11.9%	14.6%	8.3%	5.3%	3.8%	2.6%	0.7%
Qwen2.5-VL-7B (calibrated)	50.5%	16.2%	9.2%	6.7%	5.7%	4.6%	2.9%	2.8%	1.2%	0.2%
Other open general-domain VLMs										
GLM-4.6v	35.0%	14.5%	9.5%	7.9%	11.9%	7.5%	4.5%	4.2%	3.8%	1.2%
Kimi K2.5	36.4%	14.5%	9.1%	5.8%	11.1%	10.0%	7.0%	3.6%	2.2%	0.3%
Llama4 Maverick	31.8%	12.3%	9.6%	8.6%	10.3%	10.6%	7.6%	6.5%	2.5%	0.2%
Mistral Large	10.0%	13.0%	9.1%	9.8%	16.4%	15.2%	13.1%	8.0%	4.3%	1.1%
Qwen3.5 397B A17B	32.4%	12.3%	8.4%	8.4%	12.8%	11.6%	6.5%	5.2%	1.6%	0.8%
Remote-sensing / geospatial VLMs										
GeoReason	25.2%	18.3%	11.5%	11.5%	13.0%	8.0%	6.7%	3.4%	1.8%	0.6%
GeoZero	23.0%	19.6%	12.1%	11.4%	12.5%	9.8%	5.1%	3.7%	2.0%	0.8%
Proprietary general-domain VLMs										
GPT-5.2	42.6%	13.5%	10.9%	8.6%	8.1%	6.3%	4.9%	2.9%	1.7%	0.5%
GPT-5 Mini	45.2%	14.2%	10.1%	6.4%	7.3%	7.3%	4.7%	3.6%	0.8%	0.4%
GPT-5.4 Mini	35.5%	16.6%	12.3%	6.7%	8.5%	7.5%	5.4%	4.3%	2.6%	0.6%
GPT-4o Mini	36.9%	13.5%	7.2%	6.6%	10.3%	9.7%	8.6%	4.7%	2.2%	0.3%
Claude Haiku 4.5	27.6%	13.7%	11.4%	9.4%	12.7%	11.7%	4.7%	3.9%	4.0%	0.9%
Gemini 2.0 Flash Lite	55.4%	13.7%	8.0%	4.9%	6.5%	6.5%	3.0%	1.5%	0.5%	0.0%
Gemini 2.5 Flash Lite	54.8%	12.9%	7.6%	4.8%	6.7%	6.5%	3.4%	2.5%	0.7%	0.1%
Gemini 3.1 Flash	60.4%	12.2%	6.0%	4.1%	7.1%	4.5%	2.3%	3.1%	0.3%	0.0%
Grok 4 Fast	34.2%	12.0%	9.9%	6.9%	11.9%	11.8%	6.7%	4.1%	2.1%	0.4%
Reference baseline										
Random Coordinate	2.5%	6.4%	11.8%	16.7%	15.3%	14.5%	10.5%	12.3%	6.5%	3.5%

where d_{m} is the geodesic error in meters, with $\mathrm{scale} = 2327.0$ for this manifest, derived from the dataset bounds used by the evaluation script. The same run stores the sample manifest with the predictions so that the reported average score is tied to this specific 892-image evaluation set.

C Additional Results

C.1 Additional Quantitative Results

The main paper reports aggregate localization metrics, including country match, mean and median geodesic error, and GWPS. Here we provide additional distributional views of model errors. We also report controls covering direct-grid classification, calibration-stage ablation, external tools, prompt and sampling robustness, Sentinel-2 product level, and cross-domain GeoGuessr evaluation, together with qualitative examples. Implementation settings for these controls are given in Appendix B. Table 10 reports the percentage of parseable predictions falling into broad geodesic-distance bins from 0–2,000 km to errors greater than 18,000 km. These bins complement the aggregate metrics by showing whether a model's performance reflects broad regional contraction, continent-scale errors, or near-random global failures.

The distance-bin distribution reinforces the main finding that current systems often externalize coarse geographic information without reliably achieving fine localization. For example, stronger proprietary models and the calibrated Qwen2.5-VL baseline place a substantially larger fraction of predictions within 2,000 km than the random-coordinate baseline, while still leaving many predictions at multi-thousand-kilometer errors. This pattern is consistent with the view that SentinelGeo measures graded coordinate-level localization rather than a binary success/failure outcome.

Table 11: **Direct MGRS 100-km classification as a non-equivalent alternative.** City and Country are footprint-aware SentinelGeo coordinate-protocol results; MGRS Acc., Mean, and Median come from a separate direct-grid inference run on the same benchmark images. Direct-grid distances are point-based distances from the predicted grid-square center to the ground-truth coordinate and are computed only over predictions whose identifier maps to a valid grid-center coordinate; this number varies by model. The GLM-4.6v entry (mean equal to median) is therefore likely based on very few valid predictions and should not be compared with the other rows.

	Coordinate protocol		Direct-grid run		
Model	City ↑	Country ↑	MGRS Acc. ↑	Mean (km) ↓	Median (km) ↓
Qwen2.5-VL-7B (base)	0.2%	13.5%	0.10%	8879	8398
Qwen2.5-VL-7B (SFT)	2.4%	28.2%	0.10%	8853	8365
Qwen2.5-VL-7B (calibrated)	2.3%	27.3%	0.10%	8846	8335
GeoReason	0.5%	10.4%	0.10%	8906	8405
GeoZero	0.5%	10.0%	0.10%	9251	8715
Gemini 3.1 Flash	8.8%	44.4%	0.90%	4863	2295
GPT-5.2	6.5%	27.4%	0.30%	6774	5155
Qwen3.5 397B A17B	3.5%	19.8%	0.30%	11027	12291
GPT-5 Mini	5.9%	30.6%	0.20%	11382	12925
Kimi K2.5	3.7%	23.8%	0.20%	9207	9895
GPT-4o Mini	2.3%	22.6%	0.10%	9191	8689
Gemini 2.5 Flash Lite	5.8%	35.6%	0.10%	11416	13212
GLM-4.6v	2.8%	19.6%	0.10%	802	802
GPT-5.4 Mini	1.4%	18.6%	0.00%	7771	7193
Mistral Large	0.2%	3.8%	0.00%	11074	12208
Claude Haiku 4.5	2.1%	16.2%	0.00%	10875	12333
Llama4 Maverick	2.0%	20.9%	0.00%	8788	8614

C.2 Footprint-Aware Evaluation and Alternative Grid Formulations

The main and holdout tables use the footprint-aware convention defined in Appendix B.3.1.

We also tested whether direct classification over the global MGRS 100-km grid could replace the coordinate-level formulation. This is a deliberately different inference interface: the model must directly emit one of 65,230 global grid classes rather than make the coordinate-level claim used by SentinelGeo. Table 11 juxtaposes footprint-aware city/country agreement from the coordinate protocol with the error obtained by the independent direct-grid experiment. The direct MGRS square accuracy is included because exact-cell classification is the defining quantity of that alternative formulation.

Implementation details for the direct MGRS control are provided in Appendix B.3.2.

The direct-grid run only partly preserves the benchmark ranking. Across the 17 shared models, direct MGRS square accuracy has Spearman $\rho = 0.4566$ with negative median error ($p = 0.0654$) and $\rho = 0.7202$ with GWPS ($p = 0.0011$); four systems are tied at zero exact-square accuracy, and exact-square accuracy is at most 0.90% for every model. Five models also have direct-grid mean errors above 10,800 km, larger than the random-coordinate mean error in Table 1 (9831.46 km, footprint-aware), so the direct-grid interface is itself difficult for these models to use. Direct MGRS classification changes the model interface and the inference problem itself, not only the scoring view. Exact-cell accuracy also introduces hard boundaries and many ties, while geodesic distance preserves the severity of misses continuously. We therefore retain coordinate-level distance as the benchmark's primary formulation and treat direct grid classification as a non-equivalent auxiliary control rather than a substitute.

C.3 Stage-by-Stage Calibration Ablation

We compare the two calibration stages using the same 1,000-image benchmark split and the footprint-aware evaluation convention. Implementation details for the stage-wise ablation are provided in Appendix B.3.3. SFT accounts for most of the gain: relative to the base model, it reduces the

Table 12: **Stage-by-stage calibration ablation.** All rows use the same benchmark split and footprint-aware metrics; each stage is evaluated once.

Stage	N	City ↑	Country ↑	Mean (km) ↓	Median (km) ↓	GWPS ↑
Base	1000	0.2%	13.5%	6064.03	5309.93	0.0194
+ SFT	1000	2.4%	28.2%	4040.60	2157.86	0.0730
+ DAPO	1000	2.3%	27.3%	3781.32	1861.67	0.0898

Table 13: **External-tool control.** All reported localization metrics use the footprint-aware convention.

Setting	N	City ↑	Country ↑	Mean (km) ↓	Median (km) ↓	GWPS ↑
Closed-book	1000	8.8%	44.4%	3154.7	1151.1	0.2236
+ GeoGPT-RAG	1000	9.1%	37.9%	4094.6	1654.9	0.1882
+ Google Search	1000	10.0%	38.1%	4165.3	1738.2	0.1844

footprint-aware median error by 3152.07 km and increases country agreement by 14.7 percentage points. DAPO then further reduces the median by 296.19 km and the mean by 259.28 km while increasing GWPS from 0.0730 to 0.0898. City and Country agreement decrease slightly, from 2.4% to 2.3% and from 28.2% to 27.3%. In this single run, most of the gain therefore appears after SFT, and the additional DAPO change is concentrated in the distance-based metrics.

C.4 External Tools and the Agentic Setting

We test whether retrieval or an agentic search loop improves geolocation with Gemini 3.1 Flash. The closed-book condition is compared with a GeoGPT-RAG pipeline and a Google-Search-grounded agent, and both tool-augmented settings are required to query their external source. All three conditions are scored with the same footprint-aware metrics. Implementation details for the tool-augmented runs are provided in Appendix B.3.4. Neither tool-augmented configuration improves the closed-book result. GeoGPT-RAG lowers Country from 44.4% to 37.9%, increases the footprint-aware mean from 3154.7 to 4094.6 km and the median from 1151.1 to 1654.9 km, and reduces GWPS from 0.2236 to 0.1882. Google Search behaves similarly: Country falls to 38.1%, mean error rises to 4165.3 km, median error to 1738.2 km, and GWPS falls to 0.1844. City increases slightly in both tool conditions, while the country- and distance-based metrics are worse. In this setting, access to external geographic text therefore does not by itself improve localization from the overhead image. We did not test which step of the tool pipeline is responsible for the degradation.

C.5 Prompt and Sampling Robustness

We evaluate four interface conditions on the same 200-image subset: A uses the benchmark rationale-plus-token prompt, B is a semantic paraphrase, C requests coordinates directly without an externalized rationale, and D requests continuous decimal coordinates. Condition A is independently sampled three times per image. All conditions are scored with the same footprint-aware metrics. This is an exploratory comparison: each of B, C, and D is a single run, whereas the Condition-A row aggregates three runs as described below. Implementation details for prompt conditions and repeated sampling are provided in Appendix B.3.5. Across paraphrasing and coordinate representation, median errors remain in the range of hundreds to thousands of kilometers for all three models, so none of the tested interfaces approaches fine-grained localization. The direct-output condition C produces a larger degradation for the calibrated Qwen model (Country 12.5% versus 26.5–30.0% under the other conditions), whereas GPT-5 Mini and Gemini 3.1 Flash change less, so the interface effect is model dependent. Gemini 3.1 Flash remains strongest on this subset across the tested interfaces.

Repeated calls show appreciable image-level stochasticity and smaller variation in the aggregates. Across the three runs, footprint-aware median errors stay within 294 km for Qwen2.5-VL-7B (calibrated), 509 km for GPT-5 Mini, and 104 km for Gemini 3.1 Flash; City and Country agreement vary by up to 2.0 and 5.0 percentage points, respectively. At the same time, the median per-image range of geodesic errors remains large (423–1447 km), showing that individual scenes can receive

Table 14: **Prompt robustness on a shared 200-image subset.** City, Country, mean, median, and GWPS are footprint-aware. All runs use temperature 0.5. For Condition A, the three runs of each image are ordered by geodesic error and the median record is kept, so this row uses the ground truth for selection (Appendix B.3.5); B, C, and D are single runs.

Model	Cond.	N	City ↑	Country ↑	Mean (km) ↓	Median (km) ↓	GWPS ↑
Qwen2.5-VL-7B (calibrated)	A	200	1.0%	27.0%	3392.3	1886.8	0.0537
Qwen2.5-VL-7B (calibrated)	B	200	3.0%	30.0%	3664.3	1807.9	0.1051
Qwen2.5-VL-7B (calibrated)	C	200	1.0%	12.5%	4868.9	3410.2	0.0420
Qwen2.5-VL-7B (calibrated)	D	200	1.0%	26.5%	4023.8	2033.9	0.0557
GPT-5 Mini	A	200	4.0%	33.5%	4210.9	2026.9	0.1352
GPT-5 Mini	B	200	5.5%	29.5%	4048.9	2043.6	0.1182
GPT-5 Mini	C	200	4.5%	30.5%	4262.2	2331.8	0.1144
GPT-5 Mini	D	200	5.0%	29.0%	4352.7	2434.0	0.1184
Gemini 3.1 Flash	A	200	5.5%	43.5%	2828.8	1131.9	0.2005
Gemini 3.1 Flash	B	200	9.5%	47.0%	2952.7	870.6	0.2215
Gemini 3.1 Flash	C	200	10.0%	49.5%	2603.6	919.0	0.2434
Gemini 3.1 Flash	D	200	8.5%	43.0%	3359.6	1234.6	0.2106

Table 15: **Condition-A sampling consistency.** Each image is independently inferred three times; run medians, City, and Country use the footprint-aware convention. The last column is the median over images of the range (maximum minus minimum) of the three geodesic errors for that image. N = 200 for every run.

Model	Run	Median (km) ↓	City ↑	Country ↑	Median error range (km)
Qwen2.5-VL-7B (calibrated)	1	1732.3	2.0%	31.5%	1447.3
	2	1855.0	0.0%	28.0%	
	3	2026.2	2.0%	28.0%	
GPT-5 Mini	1	2068.3	4.0%	31.0%	830.3
	2	2233.7	6.0%	34.0%	
	3	2577.5	4.0%	29.0%	
Gemini 3.1 Flash	1	1007.9	8.5%	44.0%	423.5
	2	985.4	6.5%	46.5%	
	3	904.2	8.0%	49.0%	

very different predictions across runs even when the aggregates change less. We therefore treat repeated-call variation as a noise-floor diagnostic and interpret small differences conservatively.

C.6 Level-1C (TOA) versus Level-2A (BOA)

Level-1C and Level-2A are both orthorectified products; the relevant difference is atmospheric correction, from top-of-atmosphere (TOA) to bottom-of-atmosphere (BOA) reflectance. We compare the same 200 scenes using the same rendering and prompt pipeline to test whether L2A-specific radiometric processing is a confound. Both product levels are scored with the same footprint-aware metrics. The L2A rows are the Condition-A results of Table 14 on the same 200-image subset, including its median-record aggregation over three runs. Implementation details for matched L1C/L2A rendering are provided in Appendix B.3.6. The L1C condition improves the footprint-aware median for all three systems, but the magnitude and direction of the other metrics are not uniform. Qwen2.5-VL-7B (calibrated) improves in median and GWPS while its mean becomes 302.2 km worse; GPT-5 Mini improves in mean and median but its GWPS decreases slightly; Gemini 3.1 Flash improves across all reported metrics. The median shifts (98–334 km) are of the same order as repeated-call variation (Table 15), and the other metrics do not move consistently in one direction. In this small control, we therefore find no evidence that the choice of L2A changes the qualitative conclusions, although L1C gives a lower median for all three models. This remains a radiometric product-level control rather than a test of unrectified sensor geometry.

Table 16: **L1C versus L2A control.** All localization metrics use the footprint-aware convention on the same 200 matched scenes. L2A rows reproduce Condition A of Table 14.

Model	Product	N	City ↑	Country ↑	Mean (km) ↓	Median (km) ↓	GWPS ↑
Qwen2.5-VL-7B (calibrated)	L2A	200	1.0%	27.0%	3392.3	1886.8	0.0537
Qwen2.5-VL-7B (calibrated)	L1C	200	1.5%	29.0%	3694.5	1788.7	0.0839
GPT-5 Mini	L2A	200	4.0%	33.5%	4210.9	2026.9	0.1352
GPT-5 Mini	L1C	200	5.0%	35.5%	3553.8	1693.3	0.1216
Gemini 3.1 Flash	L2A	200	5.5%	43.5%	2828.8	1131.9	0.2005
Gemini 3.1 Flash	L1C	200	10.0%	49.5%	2768.3	958.5	0.2820

Table 17: **GeoGuessr-style ground-level geolocation.** We report N, Country, mean/median distance, and average task score; City is omitted because it is not provided by this evaluation. N is the manifest size. Country is normalized by all 892 images, while distances and scores are point-based and computed over non-refused predictions only (Appendix B.3.7).

Model	N	Country ↑	Mean (km) ↓	Median (km) ↓	Avg. score ↑
General-purpose API VLMs					
Gemini 3.1 Flash Lite	892	70.96%	1023.7	278.0	3771.7
Gemini 2.5 Flash Lite	892	67.38%	1389.7	311.5	3558.4
GPT-5 Mini	892	59.64%	1774.8	389.4	3342.8
Kimi K2.5	892	59.53%	1932.8	376.1	3332.7
GLM-4.6v	892	57.96%	2148.4	464.7	3202.1
GPT-5.2	892	50.00%	2313.1	528.9	3051.4
GPT-5.4 Mini	892	40.81%	2560.1	740.2	2808.9
GPT-4o Mini	892	46.64%	2863.1	686.4	2802.4
Llama4 Maverick	892	39.13%	3351.5	905.0	2536.4
Qwen3.5 397B A17B	892	39.24%	3216.0	1007.2	2478.5
Claude Haiku 4.5	892	28.92%	3835.3	1309.7	2200.0
Mistral Large	892	21.75%	4972.8	2250.7	1739.3
Local / remote-sensing-oriented models					
Qwen2.5-VL-7B (calibrated)	892	41.70%	2724.6	804.3	2718.0
Qwen2.5-VL-7B (SFT)	892	38.23%	2916.9	936.6	2564.0
Qwen2.5-VL-7B (base)	892	39.80%	3322.7	1018.4	2450.8
GeoReason	892	38.12%	3315.1	1088.1	2410.8
GeoZero	892	31.39%	4374.4	1703.7	2040.4

C.7 Cross-Domain GeoGuessr Control

To test whether the observed ordering is specific to overhead imagery, we also evaluate the same model families on 892 ground-level Google Street View images in a GeoGuessr-style task. This is intentionally a different visual domain, containing cues such as road signs, markings, vehicles, and streetscape context that are largely absent from SentinelGeo. City agreement is not computed in this evaluation. Implementation details for the street-view control and scoring are provided in Appendix B.3.7. The ground-level control shows that model behavior is domain dependent rather than following a single universal geolocation ranking. Most general-purpose API VLMs reach much lower errors on street-level images than on SentinelGeo, and their ordering differs from Table 1. Among the local models, the calibrated Qwen model is slightly better than its base model on street-level images, while the remote-sensing models GeoReason and GeoZero remain near the bottom. Overhead coordinate localization and street-view geolocation therefore rank models differently. In both settings, performance remains graded rather than binary, which is consistent with the use of metric distance rather than an exact grid label as the primary error signal.

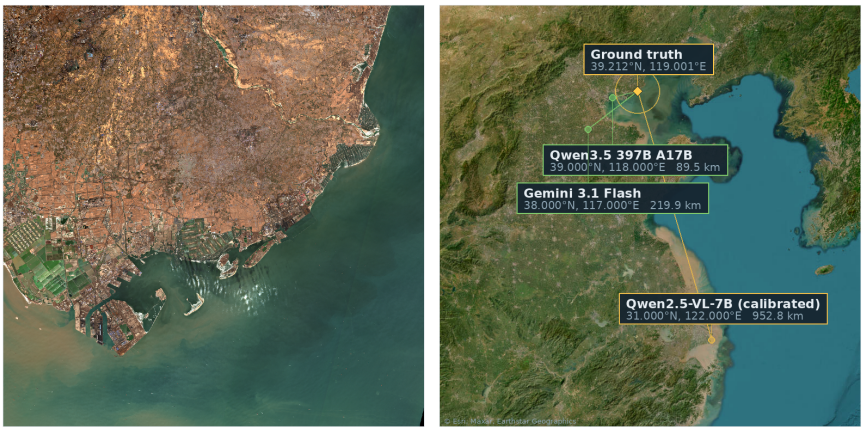
C.8 Qualitative Examples

Aggregate coordinate metrics do not fully capture whether a model's location claim is supported by the image. Figures 5–9 provide illustrative qualitative examples from SentinelGeo to complement the rationale-reliability scores reported in the main paper. The examples were selected to illustrate the

failure modes discussed in Section 4.2 and are not a random sample. Each figure shows three models; distances are raw point-based Haversine distances d_i, and long rationales are truncated for display. The examples illustrate behaviors that we observed under the image-only coordinate-token protocol: models may identify visible but geographically generic cues, hallucinate spatial evidence that is not present in the image, produce a rationale that points toward one region while emitting coordinate tokens for another, or achieve coarse regional contraction without fine localization.

These examples should be interpreted as diagnostic cases rather than a separate benchmark metric. They highlight why SentinelGeo evaluates both coordinate accuracy and rationale reliability: a prediction can be geographically closer than random while being supported by weak or post-hoc rationales, and a rationale can be visually plausible without providing enough information to narrow the location to the target region. Conversely, some failures are not purely reasoning failures; they reflect the intrinsic ambiguity of single-image overhead geolocation, where visually similar terrain, agriculture, hydrology, or settlement patterns may occur across multiple world regions.

SentinelGeo · qualitative example



Ground truth	Closest	Median	City	Country	Within footprint
39.212°N, 119.001°E	Qwen3.5 397B A17B 89.5 km	219.9 km	Bencheng	CN	0 of 3 models

Model outputs

Qwen3.5 397B A17B

The image displays a coastal region with distinct features:

1. **River Delta:** In the upper right, a meandering river flows into the sea. This matches the **Luan River** delta.
2. **Salt Pans/Aquaculture:** The land is covered in extensive rectangular grids, typical of salt evaporation ponds (Changlu Salt Field) and aquaculture farms found along the Bohai Sea coast of China.
3. **Caofeidian Island:** In the lower center, there is a distinctive curved, finger-like sandbar island. This is **Caofeidian Island**.
4. **Port/Reclamation:** To the left of the island, there is a large area of developed land and port infrastructure, which corresponds to the **Caofeidian Industrial Zone** and port, a major deep-water port project.
5. **Geography:** The coastline runs roughly northwest to southeast, facing the **Bohai Sea**. The location is the Caofeidian District of **Tangshan City, Hebei Province, China**.

The coordinates for this area are approximately 39°N latitude and 118°E longitude. The image centers roughly around 39.1°N, 118.7°E.

<LAT><N><39></LAT> **39.000°N, 118.000°E**
<LON><E><118></LON> **error 89.5 km**

Gemini 3.1 Flash

The satellite image shows a coastal region with extensive land reclamation, industrial port infrastructure, and a distinct river delta system. The geography, characterized by the specific shape of the coastline, the presence of large-scale port facilities (resembling the Tianjin Port area), and the surrounding agricultural and urban landscape, strongly points to the Bohai Sea region in China. Specifically, this is the Binhai New Area of Tianjin, located near the mouth of the Hai River. The coordinates for this region are approximately 38°N to 39°N and 117°E to 118°E.

<LAT><N><38></LAT> **38.000°N, 117.000°E**
<LON><E><117></LON> **error 219.9 km**

Qwen2.5-VL-7B (calibrated)

The scene shows a large urbanized coastal belt with a very dense patchwork of fields and built-up areas, plus a broad river system feeding into the sea. The coastline is a straight, gently curving shoreline with multiple small islands and sandbars, suggesting a major river delta meeting the ocean. The water is a muddy green near the coast, indicating sediment-laden rivers, while the open sea is deeper and bluer.

The urban fabric is extremely dense and continuous, with no mountain shadows, which points away from the Himalayas and toward the coastal plains of East Asia. The river's branching pattern and the way it spreads out into a wide estuary is characteristic of the Yangtze Delta's tributaries. The coastline here looks like the Jiangsu/Anhui coastal
... (10 more lines omitted to fit)

around 31°N and about 122°E, which is consistent with the Yangtze Delta's northern Jiangsu coast.

Focusing on the exact spot, the point should be on the coastal plain east of the river mouth, near the small islands and sandbars, matching the urban fabric and the river-fed estuary. This places the location at approximately 31.08°N, 121.75°E.

<LAT><N><31></LAT> **31.000°N, 122.000°E**
<LON><E><122></LON> **error 952.8 km**

Figure 5: **Bohai Bay coast, China.** Qwen3.5 397B A17B names specific landmarks such as Caofeidian and is closest (89.5 km); Gemini 3.1 Flash identifies the Tianjin coast (219.9 km); Qwen2.5-VL-7B (calibrated) places the scene at the Yangtze estuary (952.8 km).

SentinelGeo · qualitative example



Ground truth	Closest	Median	City	Country	Within footprint
11.254°N, 12.666°W	GPT-5.2 945.9 km	2,275 km	Lelouma	GN	0 of 3 models

Model outputs

GPT-5.2

The scene shows a broad expanse of reddish-brown lateritic soils with a dense mosaic of smallholder clearings embedded in greener savanna woodland. Drainage is dendritic with many small seasonal streams and no major river or coastal features, pointing to an interior tropical savanna/Guinea-savanna belt. The scattered pale rocky patches/inselberg-like exposures and the overall land-use texture are characteristic of inland West Africa, consistent with the southern Burkina Faso / northern Côte d'Ivoire zone. I place this in southern Burkina Faso around the 11°N band and ~4°W.

<LAT><N><11></LAT> **11.000°N, 4.000°W**
<LON><W><004></LON> **error 945.9 km**

Qwen2.5-VL-7B (calibrated)

The image shows a broad belt of dark green vegetation with a strong seasonal color contrast: broad green blocks with reddish-brown patches that look like dry-season fallow or pasture. That points to a tropical savanna or monsoonal agriculture zone rather than temperate farmland. The land is mostly cleared into irregular blocks and strips, with a few larger, more rectilinear plots, suggesting smallholder farming and shifting cultivation.

There are many small, sinuous drainage lines and low-relief ridges, but no dense mountain chains or large river floodplains. That fits the interior of West Africa rather than the Congo Basin's dense rainforest or the Sahel's more arid margins. The overall tone and land-use pattern matches the southern Benue-Trouve corridor in Nigeria—wide savanna with scattered agriculture and patchy forest.

To refine further, I look for a small, bright linear feature near the lower-left quadrant that resembles a narrow road or a small river cutting through the savanna. The surrounding land is intensively farmed and forested, consistent with the Benue State agricultural belt. The latitude band around 9°N fits the southern savanna zone of Nigeria. The longitude around 8°E places it well inside the country, not in the Atlantic coast or the far east, and the mixture of smallholder plots and forest blocks is very characteristic of Benue's interior.

Placing the point in the central-southern Benue savanna, the best match is a location near 9°N and about 8°E, where the landscape is dominated by savanna agriculture and scattered forest blocks. The specific placement aligns with the transition zone between more intensive farming to the west and more forested uplands to the east, consistent with the area around 9.1°N, 7.9°E.

<LAT><N><09></LAT> **9.000°N, 8.000°E**
<LON><E><08></LON> **error 2,275 km**

GLM-4.6v

The satellite image displays a mosaic of green vegetation and brownish earth, indicative of a tropical or subtropical region with mixed forest and savanna landscapes. The patchy, irregular vegetation pattern is characteristic of the African savanna-forest transition zone, common in Central Africa. The presence of subtle linear features (likely rivers or watercourses) and the earthy color palette (with dense greenery interspersed with drier, brownish areas) align with regions like the Democratic Republic of Congo. Given the tropical latitude and Central African longitude, the location is inferred to be around 5°N latitude and 25°E longitude, fitting the 1-degree bin format for discrete tokens.

<LAT><N><5></LAT> **5.000°N, 25.000°E**
<LON><E><25></LON> **error 4,200 km**

Figure 6: **West African savanna, Guinea.** All three models recognize a tropical savanna mosaic but name different countries: GPT-5.2 southern Burkina Faso (945.9 km), Qwen2.5-VL-7B (calibrated) central Nigeria (2,275 km), and GLM-4.6v the Democratic Republic of the Congo (4,200 km).

SentinelGeo · qualitative example



Ground truth	Closest	Median	City	Country	Within footprint
34.816°N, 2.106°E	Gemini 2.5 Flash Lite 494.7 km	2,071 km	Ksar Chellala	DZ	0 of 3 models

Model outputs

Gemini 2.5 Flash Lite

The image displays a vast arid or semi-arid landscape characterized by dry riverbeds (wadis), sparse vegetation, and a prominent salt flat or dry lake. The terrain appears to be eroded, with numerous dendritic drainage patterns. The presence of a large, bright white salt flat is a significant clue. The overall color palette of ochre, brown, and pale yellow is typical of desert and steppe regions.

Upon closer inspection of the cropped images, the patterns of erosion and the scale of the features suggest a location
… (5 more lines omitted to fit)

large endorheic salt lake in northeastern Algeria. The surrounding landscape, with its dry river systems and arid terrain, is consistent with this region.

Based on the location of Chott Melrhir, the approximate coordinates can be determined. Chott Melrhir lies in the northeastern part of Algeria.

<LAT><N><33></LAT> **33.000°N, 7.000°E**
<LON><E><007></LON> **error 494.7 km**

Qwen2.5-VL-7B

The image shows a desert landscape with sparse vegetation, a large body of water, and a few small settlements. The presence of the large body of water and the arid environment suggest a region that could be in the Middle East or North Africa. The settlements appear to be small and sparsely distributed, which is characteristic of regions like the Sahara Desert.

Given the arid and desert-like appearance, and the presence of a large body of water, it is likely this image is from a region in North Africa, possibly near the border of Egypt and Sudan, where the Nile River flows through a desert landscape.

Final geolocation tokens:
<LAT><37><N></LAT> **37.000°N, 25.000°E**
<LON><25><E></LON> **error 2,071 km**

Qwen2.5-VL-7B (calibrated)

The scene is overwhelmingly arid with tan/brown tones, sparse vegetation, and many dry drainage lines. There are no coastlines, no dense urban grid, and only a few small, bright clusters that look like salt flats or industrial/urban areas. That combination—desert interior with salt flats and a few industrial-looking patches—strongly suggests the Middle East rather than Africa or Central Asia.

The presence of a large, irregular, bright, shallow waterbody with a narrow, darker margin points to a salt lake or
… (9 more lines omitted to fit)

drainage lines feeding into it, matching the way salt lakes and dry wadis appear in the Al Ula region. The surrounding rugged ridges and the way the lake sits between them align well with the Al Ula–Yanbu–Hijaz transition zone. Placing the point on the lake's southern edge, where the shore is narrow and the surrounding terrain is rugged, lands in the Al Ula area at approximately 25.5°N and 37.6°E.

<LAT><N><26></LAT> **26.000°N, 38.000°E**
<LON><E><38></LON> **error 3,558 km**

Figure 7: **Arid steppe with a salt flat, northern Algeria.** Gemini 2.5 Flash Lite attributes the salt flat to Chott Melrhir in northeastern Algeria, which is not the lake near the target, yet is closest (494.7 km). The base Qwen2.5-VL-7B rationale points to the Egypt–Sudan border, while its tokens (written in magnitude-then-direction order) decode to 37°N, 25°E in the Aegean Sea (2,071 km). Qwen2.5-VL-7B (calibrated) places the scene near Al Ula, Saudi Arabia (3,558 km).

SentinelGeo · qualitative example



Ground truth	Closest	Median	City	Country	Within footprint
30.212°N, 54.454°E	GPT-4o Mini 49.6 km	1,096 km	Shahr-e Herat	IR	1 of 3 models

Model outputs

GPT-4o Mini

The satellite image depicts a vast arid landscape with some green patches, likely indicating agricultural areas. The topography shows a mix of mountainous regions and flat areas, characteristic of desert or semi-desert environments. The presence of numerous linear features suggests irrigation or cultivation, which can be indicative of farming practices in arid climates.

Considering these characteristics, this landscape could be located in regions with desert climates, such as parts of North Africa or the Middle East. Given the structure and layout evident in the image, it resembles areas found in Iran or parts of the Middle East like Iraq or Saudi Arabia, where irrigation is used to support agriculture amidst desert conditions.

The specific patterns of vegetation and the overall geography point towards the Iranian plateau, particularly considering the topography and agricultural development.

Thus, I conclude that the location is in Iran.

<LAT><N><30></LAT> **30.000°N, 54.000°E**
<LON><E><54></LON> **error 49.6 km**

GLM-4.6v

The satellite image depicts an arid desert landscape with sandy plains, rocky outcrops, and sparse vegetation, typical of regions like the Arabian Peninsula or the Sahara. The terrain's color palette (tan/brown) and linear features (possibly roads or irrigation) suggest a Middle Eastern desert environment. Considering discrete coordinate bins, a plausible location is in Saudi Arabia, where such desert landscapes are prevalent. The latitude (25°N) and longitude (45°E) align with the Rub' al Khali desert region.

<LAT><N><25></LAT> **25.000°N, 45.000°E**
<LON><E><045></LON> **error 1,096 km**

Qwen2.5-VL-7B (calibrated)

The scene is overwhelmingly arid with tan to light brown dunes and exposed rock, plus sparse vegetation. There are no coastal cues, no large river corridors, and no dense urban fabric, which points away from the humid Mediterranean and toward a hot desert. The terrain shows long, gently rolling dune fields with darker, more rugged ridges and small green patches, suggesting irrigated oasis-like agriculture in a desert.

The patchwork of rectangular fields is a strong clue for an irrigated oasis system, with water-intensive agriculture concentrated on narrow strips and blocks. The overall geometry—many small fields aligned in straight lines and clusters of greener vegetation—fits well with the Central Asian arid zones around the Tian Shan foothills, where
... (7 more lines omitted to fit)

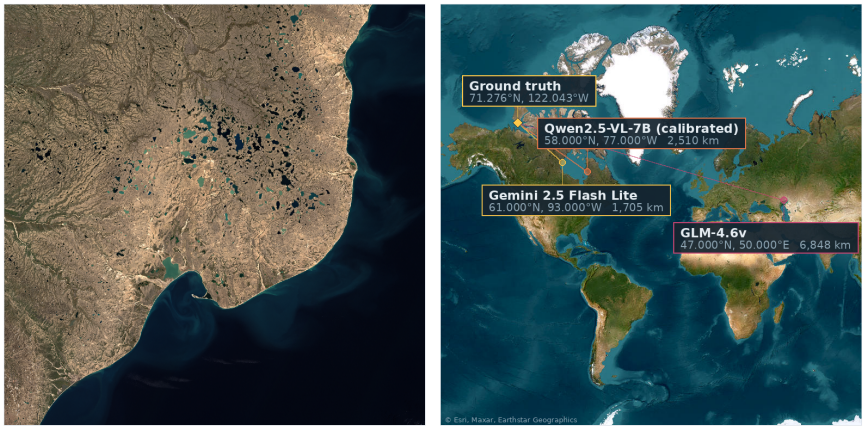
Shan near Tajikistan's desert margins. The location in this basin corresponds to a spot in the desert between agricultural strips and the nearby ridge, placing it in the mid-30s north and around 66° east.

Placing the pin at the center of the agricultural mosaic and ridge alignment in that desert basin yields the precise spot at approximately 34.8°N, 66.0°E.

<LAT><N><35></LAT> **35.000°N, 66.000°E**
<LON><E><066></LON> **error 1,204 km**

Figure 8: **Iranian Plateau.** GPT-4o Mini gives a generic arid-agriculture rationale that concludes only "Iran", yet its prediction falls within the tile footprint (49.6 km). GLM-4.6v places the scene in Saudi Arabia (1,096 km), and Qwen2.5-VL-7B (calibrated) argues for Central Asian oases but outputs 35°N, 66°E in Afghanistan (1,204 km).

SentinelGeo · qualitative example



Ground truth	Closest	Median	City	Country	Within footprint
71.276°N, 122.043°W	Gemini 2.5 Flash Lite 1,705 km	2,510 km	Norman Wells	CA	0 of 3 models

Model outputs

Gemini 2.5 Flash Lite

The image shows a coastal area with numerous small lakes scattered across a tundra-like landscape. The coastline is irregular, with a prominent bay or inlet. The water is a deep blue, suggesting a large body of water, likely an ocean or sea. The vegetation is sparse, consistent with a subarctic or arctic environment. The presence of many small, irregular lakes is characteristic of glaciated terrain, particularly in regions like northern Canada, Alaska, or Siberia.

The specific shape of the coastline and the pattern of the lakes are key to pinpointing the location. The swirl-like patterns in the water near the coast could indicate currents or sediment plumes. Observing the overall morphology, the numerous lakes and the coastal shape strongly resemble the landscape of Nunavut, Canada, particularly the
... (6 more lines omitted to fit)

suggestive of the Kivalliq Region of Nunavut, specifically areas around Chesterfield Inlet or Rankin Inlet.

After reviewing satellite imagery databases and comparing the visual cues, the area strongly matches the geography of the western coast of Hudson Bay in Nunavut, Canada. The distinct curvature of the coastline and the prevalence of numerous interconnected lakes are highly characteristic of this region.

<LAT><N><61></LAT> **61.000°N, 93.000°W**
<LON><W><93></LON> **error 1,705 km**

Qwen2.5-VL-7B (calibrated)

The scene is a broad coastal plain with a rugged, heavily glaciated interior and many small, bright-blue lakes scattered across the uplands. That combination—glacially scoured terrain, numerous lakes, and a long, smooth shoreline—fits the Canadian Shield well. The coastline is not a narrow rocky fjord; it's a wide, gently curving shore with wide beaches and a broad coastal plain, which matches the James Bay / Hudson Bay / Ungava Bay region rather than the Pacific Northwest.

The presence of a large, dark water body to the right and a long, indented shoreline to the left suggests a large bay opening onto Hudson Bay, and the overall geography points to northern Quebec. The interior shows extensive glacial
... (8 more lines omitted to fit)

Placing the point on the coastal plain near the southern Ungava Bay region, the best fit is around 58°N and about 77°W, where the terrain is glacially carved and the coastline is wide. The exact spot corresponds to the coastal plain just south of the glacially dominated interior, aligning with the observed transition from uplands to coastal plain and the bay's shape.

<LAT><N><58></LAT> **58.000°N, 77.000°W**
<LON><W><077></LON> **error 2,510 km**

GLM-4.6v

The satellite image shows a coastal region with numerous scattered lakes, arid to semi-arid terrain, and a large body of water (likely the Caspian Sea) to the east. The landscape's flatness, lake distribution, and coastal alignment match the northern Caspian Sea coast, particularly in Kazakhstan. Discrete coordinates for this region are approximately 47°N, 50°E.

<LAT><N><47></LAT> **47.000°N, 50.000°E**
<LON><E><50></LON> **error 6,848 km**

Figure 9: **Arctic coast, Northwest Territories, Canada.** Gemini 2.5 Flash Lite claims to have reviewed "satellite imagery databases" although no tools are available, and places the scene on the western Hudson Bay coast (1,705 km). Qwen2.5-VL-7B (calibrated) chooses Ungava Bay (2,510 km), and GLM-4.6v the northern Caspian coast (6,848 km).