GeoSolver: Scaling Test-Time Reasoning in Remote Sensing with Fine-Grained Process Supervision

Lang Sun[1,2], Ronghao Fu[1,2]✉, Zhuoran Duan[1,2], Haoran Liu[1,2], Xueyan Liu[1,2], and Bo Yang[1,2]✉

[1]College of Computer Science and Technology, Jilin University, Changchun 130012, China
[2]Key Laboratory of Symbolic Computation and Knowledge Engineering of Ministry of Education Jilin University
{sunlang24,duanzr24,lhr24}@mails.jlu.edu.cn,
{furh,xueyanliu,ybo}@jlu.edu.cn

Abstract. While Vision-Language Models (VLMs) have significantly advanced remote sensing interpretation, enabling them to perform complex, step-by-step reasoning remains highly challenging. Recent efforts to introduce Chain-of-Thought (CoT) reasoning to this domain have shown promise, yet ensuring the visual faithfulness of these intermediate steps remains a critical bottleneck. To address this, we introduce GeoSolver, a novel framework that transitions remote sensing reasoning toward verifiable, process-supervised reinforcement learning. We first construct GeoPRM-2M, a large-scale, token-level process supervision dataset synthesized via entropy-guided Monte Carlo Tree Search (MCTS) and targeted visual hallucination injection. Building upon this dataset, we train GeoPRM, a token-level process reward model (PRM) that provides granular faithfulness feedback. To effectively leverage these verification signals, we propose Process-Aware Tree-GRPO, a reinforcement learning algorithm that integrates tree-structured exploration with a faithfulness-weighted reward mechanism to precisely assign credit to intermediate steps. Extensive experiments demonstrate that our resulting model, GeoSolver-9B, achieves state-of-the-art performance across diverse remote sensing benchmarks. Crucially, GeoPRM unlocks robust Test-Time Scaling (TTS). Serving as a universal geospatial verifier, it seamlessly scales the performance of GeoSolver-9B and directly enhances general-purpose VLMs, highlighting its remarkable cross-model generalization. The code will be available at `https://github.com/minglangL/GeoSolver`.

Keywords: VLMs · Process Reward Model · Remote Sensing

1 Introduction

The integration of Vision-Language Models (VLMs) has profoundly advanced the field of remote sensing [15, 22, 28, 29, 41, 49, 68, 74], transitioning systems

✉ Corresponding author.

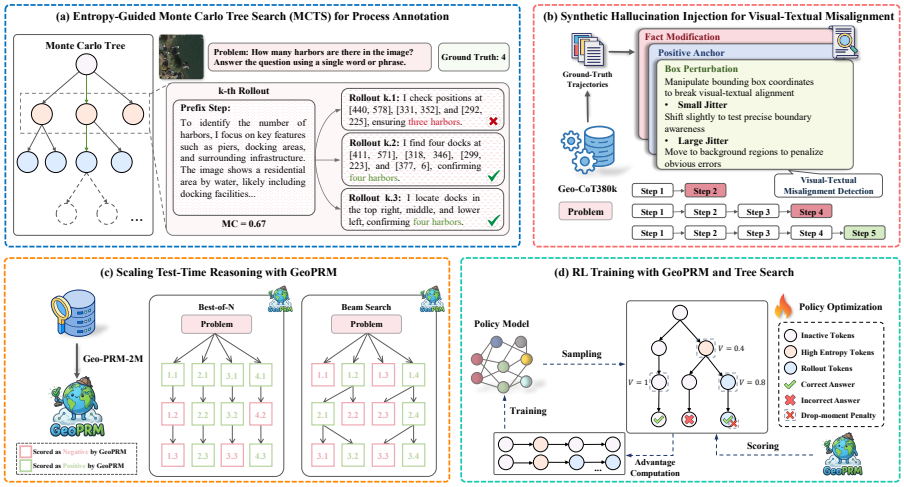


Fig. 1: Overall framework of our proposed GeoSolver. To enable reliable geospatial reasoning, we first construct the Geo-PRM-2M dataset via (a) Entropy-Guided MCTS and (b) Synthetic Hallucination Injection. The trained GeoPRM is then seamlessly integrated to (c) scale test-time reasoning via advanced search strategies during inference, and (d) align the base policy via Process-Aware Tree-GRPO during training.

from simple discriminative perception to complex, open-ended visual understanding. Recently, pioneering efforts have sought to elevate these models further by introducing the Chain-of-Thought (CoT) paradigm to the geospatial domain [12, 24, 62, 70]. By forcing models to decompose tasks and explicitly gather visual evidence before predicting an answer [30,46,53], these perceptually-grounded reasoning frameworks have demonstrated remarkable potential in interpreting dense, top-down satellite imagery.

However, optimizing these multimodal reasoning processes presents significant challenges. While traditional end-to-end remote sensing VLMs relied exclusively on Supervised Fine-Tuning (SFT) [22, 35, 45], recent reasoning-centric models have advanced by adopting Reinforcement Learning (RL) frameworks [12, 30, 53], most notably Group Relative Policy Optimization (GRPO) [10]. Yet, current GRPO paradigms in this domain rely predominantly on outcome-based rewards. In visually complex geospatial scenarios, optimizing solely on final outcomes exacerbates the credit assignment problem. Without explicit step-level verification, models are frequently rewarded for "lucky guesses," where they generate linguistically fluent but visually ungrounded claims (e.g., hallucinating an incorrect bounding box that coincidentally leads to the right object count). Consequently, outcome-supervised policies are inadvertently encouraged to memorize spurious correlations rather than perform faithful visual grounding [43, 58]. To eliminate these spatial hallucinations, transitioning to process supervision is imperative. While Process Reward Models (PRMs) have shown tremendous promise in domains like mathematical reasoning [19, 60, 63, 76], their application

in remote sensing remains entirely uncharted. Furthermore, naively importing scalar PRMs as direct optimization targets is highly problematic, as it frequently induces length bias and reward hacking, driving the policy to generate artificially truncated reasoning to avoid step-wise penalties [37, 42].

To overcome these fundamental challenges, we present the **GeoSolver** framework, a novel approach that aligns remote sensing VLMs towards verifiable, step-by-step reasoning. Recognizing that robust process supervision requires domain-specific discriminative signals, we first construct **Geo-PRM-2M**, a large-scale token-level process supervision dataset. We synthesize this corpus through a dual-view pipeline: we employ an entropy-guided Monte Carlo Tree Search [4] to mine the model's intrinsic logical errors and inject synthetic perturbations (e.g., bounding box jittering) to explicitly target visual grounding failures. Trained on this dataset, our token-level verifier, **GeoPRM**, provides granular, continuous feedback on reasoning faithfulness.

Building upon GeoPRM, we propose a novel RL alignment algorithm: **Process-Aware Tree-GRPO**. Rather than relying on sample-inefficient linear rollouts [47], we construct an entropy-guided reasoning tree during the RL exploration phase. Crucially, we introduce a faithfulness-weighted reward mechanism that penalizes trajectories exhibiting sudden drops in PRM confidence [37], even if their final outcome is correct. By propagating these process-aware signals through the tree via global and local advantages, our algorithm explicitly incentivizes reasoning paths that are not only accurate but consistently verifiable.

Extensive experiments across six major remote sensing tasks demonstrate the efficacy of our approach. Applying our framework to train the GeoSolver-9B model, we show that under standard inference, it significantly outperforms both dedicated remote sensing models and general-purpose reasoning VLMs. Furthermore, by employing GeoPRM for test-time scaling (TTS) [38, 66], we observe consistent performance gains, validating the compute-optimal scaling law in remote sensing. Most notably, when used to guide general-purpose models, GeoPRM enables them to surpass fully fine-tuned domain experts, demonstrating that our reward model captures a generalized, fundamental logic of multimodal geospatial verification.

In summary, our contributions can be summarized as follows:

- We construct Geo-PRM-2M, the first large-scale process supervision dataset for remote sensing, and develop GeoPRM, a token-level PRM that precisely localizes logical and visual hallucinations.
- We propose Process-Aware Tree-GRPO, an RL algorithm that organically integrates efficient tree-structured exploration with step-wise verification, resolving the credit assignment inherent in standard reasoning alignment.
- Experimental results show that GeoSolver-9B achieves leading performance across diverse geospatial tasks. Furthermore, we comprehensively validate the efficacy of Test-Time Scaling in remote sensing and showcase the remarkable cross-model generalization of our proposed PRM.

2 Related Work

Vision-Language Models in Remote Sensing. While generalist Vision-Language Models (VLMs) like GLM-4.1V [52] and Qwen3-VL [2] exhibit remarkable capabilities, they often struggle with the unique top-down perspectives and scale variations of remote sensing imagery [35, 67, 74]. Domain-specific models (e.g., GeoChat [22], VHM [41]) address this domain gap but typically operate as black boxes, lacking interpretable reasoning. To introduce transparency, pioneering frameworks like RS-EoT [46] and GeoZero [53] integrate Chain-of-Thought (CoT) [54] for step-by-step geospatial analysis. However, as these models transition from Supervised Fine-Tuning (SFT) to Reinforcement Learning (RL), they predominantly rely on outcome-based rewards. Without explicit step-level verification, they are prone to visual hallucinations, often being rewarded for flawed intermediate logic that coincidentally yields the correct final answer [18, 30, 62]. This highlights the critical need for process-supervised alignment in remote sensing.

Process Reward Models. Unlike Outcome Reward Models (ORMs) [8] that evaluate only the final answer, Process Reward Models (PRMs) [27, 73] provide granular feedback by assessing each intermediate reasoning step. To bypass the high costs of human annotation seen in early PRMs, recent natural language and math frameworks (e.g., URSA [37], GraphPRM [42]) employ Monte Carlo Tree Search (MCTS) [4] for automated data synthesis. However, multimodal PRMs remain largely underexplored, particularly within the remote sensing domain. The geospatial field introduces unique challenges where errors frequently arise from subtle visual-textual misalignments rather than purely logical flaws [23, 40]. GeoPRM bridges this gap by utilizing the Geo-PRM-2M dataset, which is constructed through Automated Process Annotation via Entropy-Guided MCTS and Synthetic Hallucination Injection, ensuring robust token-level verification.

Reinforcement Learning with Tree Search. RL, particularly policy optimization algorithms like PPO [44] and GRPO [10], has become the standard for aligning VLMs. While integrating process supervision into RL offers clear advantages over outcome supervision [26,73], directly using static scalar PRM scores as optimization targets often induces reward hacking and length bias, causing models to generate artificially truncated responses [37, 42]. Concurrently, while tree search algorithms [16, 20] are widely used for Test-Time Scaling (TTS), their integration into online RL training is computationally prohibitive. Frameworks like TreeRL [17] mitigate this via entropy-guided exploration during rollout. Building on these insights, we propose Process-Aware Tree-GRPO. To prevent reward hacking, we replace scalar score accumulations with a drop-moment penalty, and we systematically propagate these robust verification signals through an entropy-guided reasoning tree to optimize the policy efficiently.

3 Methodology

In this section, we introduce the proposed GeoSolver in detail. We first provide the preliminaries of the base model and Supervised Fine-Tuning (SFT) initial-

ization in Sec. 3.1. Then, to address the challenge of granular verification, we clarify the formulation of our token-level Process Reward Model (GeoPRM) in Sec. 3.2 and the Process-Aware Tree-GRPO alignment algorithm in Sec. 3.3.

3.1 Preliminaries

Base Vision-Language Model. We build our framework upon the pretrained GLM-4.1V-9B-Base [52]. This state-of-the-art foundation model features a decoupled architecture ideally suited for geospatial tasks. For visual encoding, it employs Aimv2-Huge [13], a vision transformer capable of handling variable image resolutions and aspect ratios as a critical requirement for processing multi-scale remote sensing imagery. Unlike traditional fixed-resolution encoders, Aimv2-Huge utilizes a dynamic positional encoding scheme that adapts its pretrained position table via bicubic interpolation, allowing it to preserve fine-grained spatial details without distortion. For language modeling, the model incorporates a 3D-RoPE decoder, which further enhances its ability to comprehend complex spatial relationships within the visual context.

Supervised Fine-Tuning. To equip the base model with the fundamental capability for Chain-of-Thought (CoT) reasoning, we perform supervised fine-tuning on the Geo-CoT380k dataset [30]. This dataset contains 380k samples annotated with structured reasoning paths (e.g., *planning* $\rightarrow$ *visual evidence collection* $\rightarrow$ *synthesis*). Given an input image I, a query Q, and the corresponding ground-truth reasoning chain $C = \{c_1, c_2, \ldots, c_{|C|}\}$, we train the model to minimize the negative log-likelihood of the target tokens:

$$\mathcal{L}_{\mathrm{SFT}}(\theta) = -\mathbb{E}_{(I,Q,C)\sim\mathcal{D}_{\mathrm{sft}}} \sum_{t=1}^{|C|} \log \pi_\theta(c_t | c_{<t}, I, Q) \tag{1}$$

We denote the policy derived from this stage as π_{sft}. While π_{sft} successfully adopts the desired reasoning format, it remains susceptible to logical errors and visual hallucinations, serving as the starting policy for our subsequent reinforcement learning alignment.

3.2 Process Reward Modeling

To enable granular verification of reasoning trajectories, we develop GeoPRM, a token-level discriminator designed to assign a scalar likelihood of correctness to each generated token. Unlike step-level reward models, our token-level architecture allows for precise localization of errors such as incorrect bounding box coordinates within the sequence. Based on Geo-CoT380k, we construct GeoPRM-2M, a composite dataset of approximately 2 million samples, through a dual-strategy pipeline.

Automated Process Annotation via Entropy-Guided MCTS. To automate process annotation, recent works in reasoning domains typically rely on Monte Carlo (MC) method [36,37,42]. However, applying uniform random sampling to the vast multimodal solution space often yields redundant paths that fail

to capture the critical divergence points between successful and failed reasoning. To efficiently explore the solution space and establish clear decision boundaries, we design an Entropy-Guided Monte Carlo Tree Search algorithm. This approach prioritizes exploration in regions of high uncertainty within the policy π_{sft}.

Tree Construction. We iteratively build the reasoning tree by identifying forking points where the model is uncertain. Specifically, we compute the entropy $H(y_{\text{next}}|I, Q, y_{<t})$ of the next token distribution. We select the top-N tokens with the highest entropy as branching nodes and expand them by performing T rollouts. This process iterates for K rounds, constructing a dense tree of diverse reasoning paths encompassing both correct and flawed logic.

Value Estimation and Labeling. Once the tree is constructed, we estimate the faithfulness of each step s_t based on its rollout success rate. Formally, given the image I, query Q, and a reasoning prefix $s_{<t}$, the MC value $V(s_t)$ is defined as the proportion of its T rollouts leading to the correct final answer y_{gt}:

$$V(s_t) = \frac{1}{T} \sum_{k=1}^{T} \mathbb{I}(o_k = y_{\text{gt}}|s_{<t}, I, Q) \tag{2}$$

Using this strategy, we generate a pool of 3.72 million reasoning trajectories. To ensure the dataset contributes valid discriminative signals, we calculate the standard deviation of correctness scores within each problem's solution space. We filter out problems with low variance (where the model is either uniformly correct or uniformly incorrect across all paths), retaining 1.37 million high-value samples that effectively highlight the boundary between valid reasoning and plausible errors.

Synthetic Hallucination Injection. Apart from logical errors, perceptual inconsistency between images and text is a unique and prominent challenge in multimodal scenarios [14, 59]. While MCTS-based annotation captures logical uncertainty, it may fail to isolate subtle visual-textual misalignments if the model hallucinates its way to the right answer. To explicitly penalize ungrounded visual claims, we propose a targeted synthetic injection engine to construct a subset derived from ground-truth trajectories. We employ two perturbation strategies: 1) Box Perturbation: We manipulate bounding box coordinates to break visual-textual alignment, creating *Small Jitter* (shifting slightly to test precise boundary awareness) and *Large Jitter* (moving to background regions to penalize obvious errors). 2) Fact Modification: We alter factual statements, such as object counts or attributes, creating *Tampered* negatives that violate the visual premise.

Combined with unaltered positive anchors from Geo-CoT380k, this synthetic engine contributes approximately 0.7 million samples, bringing the total GeoPRM-2M training instances to 2 million.

Training Objective. We initialize GeoPRM from π_{sft} and append a linear binary classification head. To evaluate the current token within the full reasoning context, we concatenate the image I and query Q with all preceding tokens. Formally, we define the training as a dense token-level binary classification task

over a sequence of length L:

$$\min_{\theta} \mathcal{L}(\theta) = -\frac{1}{L} \sum_{t=1}^{L} m_t \left[y_t \log \hat{y}_t + (1 - y_t) \log(1 - \hat{y}_t) \right] \tag{3}$$

where y_t is the label for the t-th token, $\hat{y}_t = P(y_t | I, Q, y_{<t})$ is the predicted probability, and the mask m_t restricts the loss calculation to valid reasoning tokens.

3.3 Process-Aware Tree-GRPO

Vanilla GRPO normalizes in-group rewards based solely on final outcomes. This presents two critical limitations: 1) Reward Sparsity and Hacking: Outcome rewards ignore intermediate quality. Conversely, directly using scalar PRM scores as optimization targets induces a length bias. Because PRMs are trained on distributions where longer reasoning chains inherently carry a higher statistical risk of accumulating errors, they naturally tend to assign lower scores to longer sequences. 2) Inefficient Exploration: Independent linear sampling fails to leverage the hierarchical and compositional nature of spatial reasoning. To address these limitations, we present Process-Aware Tree-GRPO, integrating entropy-guided tree search with a drop-moment penalty mechanism, as illustrated in Figure 2.

Process-Aware Reward via Drop-Moments. To extract reliable verification signals from GeoPRM, we adopt the concept of the drop-moment [37], signifying a sudden loss of confidence between consecutive steps. For a trajectory o_i with a GeoPRM reward sequence $\{r_1^i, r_2^i, \ldots, r_{L_i}^i\}$, a drop-moment occurs if:

$$\delta_p^i = \max_j \left(r_{j-1}^i - r_j^i \right) > \rho \tag{4}$$

where ρ is the sensitivity threshold. In the remote sensing domain, the final outcome of a trajectory is typically evaluated by a continuous domain-specific scoring mechanism, yielding an outcome score $S_{\text{out}}(o_i) \in [0, 1]$ (e.g., IoU or mAP). When a drop-moment is detected, we apply a penalty factor $\gamma \in (0, 1)$ to this base score:

$$R = \begin{cases} S_{\text{out}}(o_i), & \text{if } \delta_p^i < \rho \\ \gamma \cdot S_{\text{out}}(o_i), & \text{if } \delta_p^i \geq \rho \end{cases} \tag{5}$$

This formulation circumvents the inherent length bias by focusing on relative confidence drops rather than absolute score accumulations, ensuring the policy learns from trajectories that are both outcome-accurate and process-rigorous.

Unifying Tree Search with Process Advantages. Although the drop-moment mechanism effectively refines the trajectory reward $R(o_i)$ based on process quality, this scalar remains a sequence-level reward. To explicitly assign credit to intermediate reasoning tokens, we must distribute this reward along the reasoning path. Therefore, during the rollout stage, we employ the aforementioned entropy-guided tree construction to yield a hierarchical reasoning tree $\mathcal{T}$ for a problem (see the left of Figure 2).

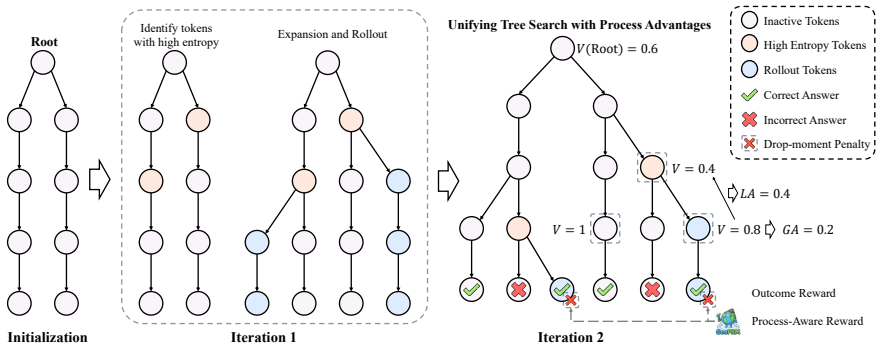


Fig. 2: Illustration of the Process-Aware Tree-GRPO. The reasoning tree dynamically expands by identifying high-entropy tokens, followed by trajectory rollouts. Leaf nodes receive outcome rewards which are further refined by GeoPRM's drop-moment penalty. These process-aware signals are then aggregated upwards to compute Local Advantage (LA) and Global Advantage (GA) for fine-grained policy optimization.

For any intermediate step s_n, let $L(s_n)$ denote the set of all complete leaf trajectories descending from it. The value of step s_n is computed as the average process-aware reward of its descendants:

$$V(s_n) = \frac{1}{|L(s_n)|} \sum_{l \in L(s_n)} R(l) \tag{6}$$

By aggregating the drop-moment penalized rewards from the leaves, $V(s_n)$ accurately reflects the potential of step s_n to lead to a faithful and correct conclusion. Based on this value, we define the Global Advantage as $GA(s_n) = V(s_n) - V(\text{root})$, measuring the node's performance relative to the overall tree expectation; and the Local Advantage as $LA(s_n) = V(s_n) - V(p(s_n))$, measuring the improvement over its immediate parent $p(s_n)$. The final re-weighted advantage is:

$$\hat{A}(s_n) = \frac{GA(s_n) + LA(s_n)}{\sqrt{|L(s_n)|}} \tag{7}$$

Because non-leaf steps appear in multiple complete sequences and would be repeatedly computed during optimization, we downweight their advantage by dividing by $\sqrt{|L(s_n)|}$, the square root of the number of leaf nodes in the subtree. This critical adjustment prevents the policy from overfitting to early, shared reasoning stages.

Finally, we update the policy π_θ using the clipped surrogate objective over all nodes in the tree:

$$\mathcal{L}_{\text{Tree-GRPO}}(\theta) = -\mathbb{E}_{(I,Q)\sim\mathcal{D},\mathcal{T}\sim\pi_{\theta_{old}}} \frac{1}{|\mathcal{T}|} \sum_{s_n \in \mathcal{T}} \min\left(r_\theta(s_n)\hat{A}(s_n), \text{clip}\left(r_\theta(s_n), 1-\epsilon, 1+\epsilon\right)\hat{A}(s_n) \right) \tag{8}$$

Table 1: Comparison of GeoSolver with state-of-the-art models on object-level remote sensing tasks. We report mIoU and mAP@50 for Visual Grounding (DIOR-RSVG [65], RRSIS-D [31], VRSBench [25], and RSVG [50]) and Object Detection (DOTAv2-val [11, 21, 55] and HRRSD [69]), alongside Accuracy for Object Counting (RSOD [33], NWPU-VHR [6], DOTAv2-val [11, 21, 55], and HRRSD [69]). The best results are bolded, and the second-best are underlined.

Method	Visual Grounding				Object Detection		Object Counting			
	DIOR	RRSIS	VRS	RSVG	DOTA	HRRSD	RSOD	VHR	DOTA	HRRSD
Closed-source Commercial Vision-Language Models										
Claude-Sonnet-4 [1]	8.87	11.62	6.65	2.53	3.89	14.87	51.5	25.0	25.17	50.11
Gemini-2.0-Flash [9]	11.40	16.05	13.52	1.75	14.30	28.92	**63.5**	39.0	29.36	54.65
GPT-5 [39]	8.27	10.60	5.65	1.49	8.66	13.15	40.0	58.0	<u>36.20</u>	58.50
Open-source Reasoning Vision-Language Models										
Kimi-VL-Thinking [51]	–	–	–	–	–	–	15.5	53.0	30.68	46.26
GLM-4.1V-Thinking [52]	39.41	43.03	<u>43.20</u>	<u>19.35</u>	<u>40.45</u>	<u>55.53</u>	28.5	<u>62.5</u>	29.80	58.96
RS-EoT [46]	9.23	12.09	1.28	11.79	–	–	35.0	49.5	28.04	57.62
Open-source Remote Sensing Vision-Language Models										
GeoChat [22]	29.87	30.30	23.66	0.13	–	–	23.5	47.0	29.87	43.22
VHM [41]	<u>49.90</u>	<u>55.20</u>	34.91	5.80	2.37	12.47	16.0	48.5	32.67	46.71
SkySenseGPT [35]	27.76	33.63	21.25	4.57	4.56	6.23	<u>51.5</u>	49.5	33.11	58.73
EarthDial [49]	39.46	35.39	13.04	7.18	3.52	8.05	41.0	52.5	32.23	<u>61.45</u>
GeoSolver	**75.62**	**76.66**	**64.57**	**27.43**	**53.47**	**94.74**	45.5	**79.0**	**45.92**	**84.13**

where $r_\theta(s_n) = \frac{\pi_\theta(s_n)}{\pi_{\theta_{old}}(s_n)}$ denotes the probability ratio. By explicitly linking drop-moments to leaf rewards and propagating them via local and global advantages through the tree, this algorithm systematically aligns the model toward verifiable, hallucination-free reasoning.

4 Experiments

4.1 Experimental Setup

Datasets and Tasks. To comprehensively validate the versatility of GeoSolver, we evaluate its performance across six major remote sensing tasks: Object Counting (OC), Object Detection (OD), Visual Grounding (VG), Scene Classification (SC), Visual Question Answering(VQA) and Image Captioning(IC). These tasks are assessed on a diverse suite of 17 established benchmark datasets. For training, we utilize the Geo-CoT380k dataset [30] for the initial SFT stage, and its expanded training set for the Tree-GRPO alignment.

Baselines. We benchmark GeoSolver against a broad spectrum of state-of-the-art models, categorized into three groups: (1) leading closed-source commercial systems (e.g., Gemini-2.0-Flash [9], Claude-Sonnet-4 [1]); (2) open-weight general and domain-specific VLMs (e.g., Qwen2.5-VL [3], GeoChat [22], EarthDial [49]); and (3) recent reasoning-centric or CoT-enabled frameworks (e.g.,

Table 2: Comparison of GeoSolver with state-of-the-art models on scene-level tasks. We report Accuracy for Scene Classification (NWPU-RESISC45 [5], AID [56], WHU-RS19 [57], SIRI-WHU [71, 72, 75], and UCMerced [61]) and VQA (VRSBench [25] and RSVQA-HR [32]). For Image Captioning (RSITMD [64], NWPU-Captions [7], and RSICD [34]), we report the BLEU-4 metric.

Method	Scene Classification					VQA		Image Caption		
	NWPU	AID	RS19	SIRI	UCM	VRS	HR	RSIT	NWPU	RSIC
Closed-source Commercial Vision-Language Models										
Claude-Sonnet-4	58.44	60.33	76.32	64.33	67.86	58.95	55.94	20.14	28.32	11.58
Gemini-2.0-Flash	74.89	76.00	90.00	72.00	85.95	61.92	49.95	15.73	20.55	10.85
GPT-5	82.22	75.50	<u>95.53</u>	75.00	88.57	62.20	65.94	27.27	39.62	16.83
Open-source Vision-Language Models										
MiniGPT-v2	32.67	27.17	30.79	26.67	32.86	37.35	50.95	25.45	37.75	15.40
Qwen2.5-VL	68.89	71.67	86.05	67.33	78.33	59.64	57.43	27.92	38.89	17.80
Open-source Reasoning Vision-Language Models										
Kimi-VL-Thinking	72.22	70.50	88.68	69.00	77.62	<u>69.09</u>	70.93	24.82	34.84	15.60
GLM-4.1V-Thinking	70.09	69.67	86.84	60.33	82.86	63.57	55.94	20.57	29.59	12.57
RS-EoT	75.56	75.83	90.25	**78.88**	85.95	69.00	70.05	25.16	37.27	23.02
Open-source Remote Sensing Vision-Language Models										
VHM	<u>91.33</u>	<u>79.00</u>	91.84	64.33	<u>89.29</u>	54.26	69.43	38.93	50.69	25.66
SkySenseGPT	83.33	75.50	93.16	55.33	85.00	48.59	63.44	37.76	23.33	**42.47**
EarthDial	76.67	67.33	88.76	73.42	80.71	39.18	<u>72.43</u>	<u>42.09</u>	<u>67.14</u>	29.09
GeoSolver	**96.44**	**98.33**	**99.50**	<u>76.00</u>	**92.38**	**77.02**	**73.19**	**52.50**	**80.93**	<u>36.18</u>

GLM-4.1V-9B-Thinking [52], RS-EoT-7B [46]). Furthermore, for Test-Time Scaling (TTS) evaluations, we compare GeoPRM against existing general mathematical PRMs and general VLMs acting as verifiers.

Implementation Details. GeoSolver is initialized from the GLM-4.1V-9B-Base [52] checkpoint. All training stages are conducted on a cluster of four NVIDIA H200 GPUs. Specifically, GeoPRM is trained on our Geo-PRM-2M dataset for 2 epochs with a batch size of 128. The policy model undergoes 1 epoch of SFT, followed by 1000 optimization steps of Process-Aware Tree-GRPO.

4.2 Comparison with State-of-the-Art

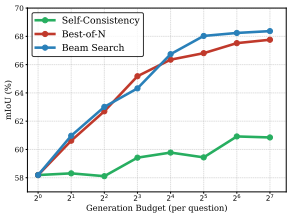
We first evaluate the standard inference performance of GeoSolver. Table 1 and Table 2 present the quantitative results across object-level and scene-level tasks, respectively. Despite the absence of inference-time search, GeoSolver consistently outperforms existing domain-specific models like GeoChat [22] and VHM [41]. On fine-grained tasks such as VG and OD, GeoSolver achieves clear performance margins over general-purpose reasoning models (e.g., GLM-4.1V-Thinking [52]). This demonstrates that the Process-Aware Tree-GRPO training endows the policy with a strong intrinsic capability for faithful spatial reasoning, effectively curbing object hallucinations during standard autoregressive generation.

Table 3: Comparison of different verification strategies on GeoSolver. We evaluate greedy decoding (GeoSolver w/o TTS), Self-Consistency (majority voting), and our GeoPRM utilizing both Best-of-N and Beam Search strategies. The generation budget is set to 32.

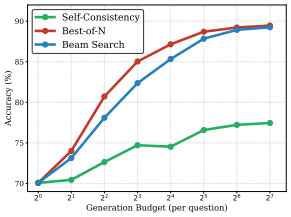
Method	VG	Detect	OC	SC	VQA	IC	↑ (Avg)
GeoSolver (w/o TTS)	58.19	74.11	54.06	90.62	70.07	47.00	–
Self-Consistency	59.44	75.17	59.28	92.57	76.58	47.51	4.1%
Best-of-N	66.35	82.51	73.92	96.01	**88.70**	**48.18**	15.8%
Beam Search	**68.04**	**84.66**	**75.45**	**98.39**	87.84	48.05	**17.5%**

Table 4: Cross-model reward model comparison using BoN performance ($N = 32$). We evaluate the verification efficacy of various strategies, including Self-Consistency, generic VLMs acting as verifiers, open-source PRMs, and our domain-specific GeoPRM across diverse remote sensing tasks.

Method	VG	Detect	OC	SC	VQA	IC	Avg
Self-Consistency	59.44	75.17	59.28	92.57	76.58	47.51	68.43
Open-source Vision-Language Models							
Qwen3-VL-8B	54.20	75.25	50.23	91.09	70.45	47.72	64.82
GLM-4.1V	53.01	74.62	47.36	90.78	71.85	47.24	64.14
Open-source Reward Models							
GraphPRM	53.82	72.92	42.36	89.58	68.35	47.17	62.37
URSA-8B-RM	53.27	73.30	45.15	86.43	70.50	46.98	62.61
GeoPRM	**66.35**	**82.51**	**73.92**	**96.01**	**88.70**	**48.18**	**75.95**



(a) Visual Grounding (VG)



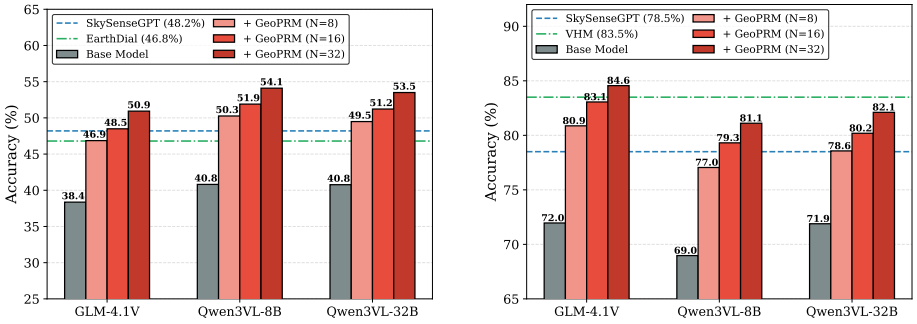
(b) VQA

Fig. 3: Compute-optimal scaling behavior of GeoSolver equipped with GeoPRM. The curves illustrate the performance gains on (a) VG and (b) VQA as the compute budget (N) increases.

4.3 Scaling Test-Time Reasoning with GeoPRM

To investigate the potential of compute-optimal inference in remote sensing, we comprehensively evaluate Test-Time Scaling (TTS) enabled by our process reward model, GeoPRM. We design three sets of experiments to assess the efficacy of different search strategies [48], the scaling behavior with increased compute budgets, and the superiority of our domain-specific PRM over generic verifiers. Note that for all aggregated task columns in this section, the reported values represent the average performance across their corresponding datasets.

Efficacy of Verification Strategies. We first evaluate the impact of different test-time verification methods applied directly to our GeoSolver policy. As shown in Table 3, relying solely on greedy decoding (w/o TTS) yields a solid baseline, but lacks a mechanism to correct intermediate logical flaws. While Self-Consistency (majority voting) provides marginal improvements, it is fundamentally bottlenecked by the base policy's distribution. In contrast, integrating GeoPRM as the verifier, unlocking explicit process supervision, leads to substantial gains. Both Best-of-N and Beam Search strategies consistently elevate performance across all metrics, with Beam Search demonstrating particular efficacy in densely grounded tasks like Object Counting and Detection by pruning hallucinatory branches early in the reasoning tree.



(a) Object Counting (OC) **(b)** Scene Classification (SC)

Fig. 4: Cross-model generalization capabilities of GeoPRM using BoN performance. The gray bars indicate the base model performance, while the red gradient bars show consistent accuracy gains as the generation budget (N) increases. Horizontal dashed lines represent the baseline performance of fully fine-tuned remote sensing models. Notably, with $N = 32$, GeoPRM-guided generalists surpass these domain experts.

Compute-Optimal Scaling Behavior. A hallmark of effective test-time reasoning is predictable performance scaling concerning the inference compute budget. We vary the generation budget to observe this scaling law. As illustrated in Figure 3, scaling the test-time budget consistently and monotonically improves performance. Notably, on complex reasoning tasks such as VG and VQA, the performance curves exhibit a robust log-linear improvement before plateauing, confirming that GeoPRM effectively harnesses additional compute to navigate complex geospatial logic without suffering from reward over-optimization.

Cross-Model Reward Model Comparison. A critical determinant of TTS efficacy is the domain alignment of the reward model itself. In Table 4, we benchmark GeoPRM against alternative verification systems, including general-purpose VLMs acting as verifiers and state-of-the-art out-of-domain mathematical PRMs like URSA-8B-RM [37]. The results reveal a stark contrast: general-domain PRMs and standard VLMs struggle to accurately assess geospatial coordinate logic and dense visual relations, often underperforming simple Self-Consistency. Conversely, GeoPRM provides highly calibrated, token-level feedback for spatial relations and visual evidence, empirically validating the absolute necessity of our domain-specific Geo-PRM-2M construction pipeline.

4.4 Cross-Model Generalization

A compelling indicator of a robust PRM is its ability to generalize beyond the specific policy it was trained alongside. In this section, we investigate whether GeoPRM can function as a universal verifier by scoring the inference rollouts of general-purpose VLMs. Specifically, we apply GeoPRM to guide GLM-4.1V [52], Qwen3-VL-8B, and Qwen3-VL-32B [2]. As illustrated in Figure 4, equipping these generalist base models with GeoPRM via TTS ($N \in \{8, 16, 32\}$) yields sub-

Table 5: Ablation study of different RL alignment strategies on GeoSolver. We evaluate the progressive integration of Vanilla GRPO, Average Process Score (APS), our Process-Aware (PA), and Tree-based exploration.

Models	Tree	PRM		VG	OC	Det	IC	SC	VQA	Avg
		APS	PA	(mIoU)	(Acc)	(mAP)	(BLEU-4)	(Acc)	(Acc)	
Base (GLM-4.1V-Base)	–	–	–	34.48	16.96	3.56	12.43	59.78	8.16	22.56
+ SFT	–	–	–	53.77	49.73	59.42	44.90	86.01	68.23	60.34
Vanilla GRPO	✗	✗	✗	59.31	55.56	68.89	54.89	91.74	72.17	67.09
+ Average Process Score	✗	✓	✗	48.69	48.22	50.76	46.45	87.21	67.44	58.13
+ Process-Aware	✗	✗	✓	<u>60.12</u>	<u>59.21</u>	69.90	54.67	<u>92.01</u>	73.02	<u>68.16</u>
+ Tree-based	✓	✗	✗	59.46	57.77	<u>70.82</u>	<u>55.21</u>	91.83	<u>73.68</u>	68.13
Process-Aware Tree-GRPO	✓	✗	✓	**61.07**	**63.6**	**73.97**	**56.54**	**92.73**	**75.11**	**70.51**

Table 6: Ablation of Geo-PRM-2M training data components evaluated using BoN performance. We compare reward models trained on the full dataset against those trained without intrinsic MCTS samples (w/o S_{MC}) or without Synthetic Hallucination Injection (w/o S_{SHI}).

Models	Dataset	Object-level				Scene-level			
		N=4	N=8	N=16	N=32	N=4	N=8	N=16	N=32
GeoSolver	Geo-PRM-2M	**68.44**	**70.92**	**73.26**	**74.26**	74.03	**75.87**	**76.97**	**77.63**
	w/o S_{MC}	62.74	63.60	63.82	64.27	69.53	70.21	69.82	70.77
	w/o S_{SHI}	66.28	67.81	69.72	70.26	**74.24**	75.46	76.37	76.89
Qwen3-VL-8B	Geo-PRM-2M	**28.70**	**30.98**	**32.56**	**34.13**	**48.51**	**50.45**	**51.93**	**52.99**
	w/o S_{MC}	24.87	25.38	25.22	26.11	45.11	45.88	46.04	45.97
	w/o S_{SHI}	26.85	27.88	28.30	28.92	47.80	49.57	50.24	50.78

stantial and consistent performance improvements over their greedy decoding. More importantly, with a sufficient compute budget (e.g., $N = 32$), these general-purpose models successfully surpass the performance of fully fine-tuned, domain-specific remote sensing models (such as SkySenseGPT [35], EarthDial [49], and VHM [41]), which are denoted by the horizontal reference lines. This cross-model improvement underscores that GeoPRM has not merely overfit to GeoSolver's output distribution; rather, it has successfully internalized a generalized, transferable discriminative logic for multimodal geospatial verification.

4.5 Ablation Study

To systematically validate our proposed components, we ablate the RL alignment strategies and the PRM training data composition.

Reward Formulation and Tree Exploration. In Table 5, we systematically ablate the components of our RL alignment phase. First, compared to Vanilla GRPO (only outcome rewards), integrating a naive Average Process Score (APS) induces severe reward hacking: the policy artificially truncates reasoning chains

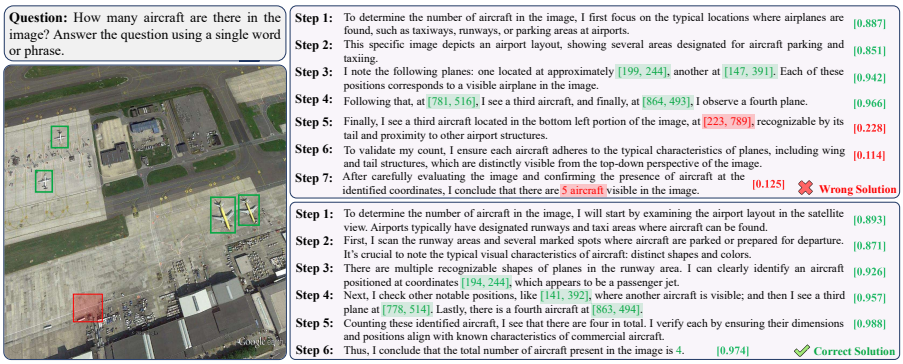


Fig. 5: Qualitative example of GeoPRM verification. We visualize two reasoning trajectories for an aircraft counting task. In the top erroneous attempt, GeoSolver hallucinates a non-existent aircraft at Step 5, which GeoPRM instantly penalizes with a sharp score drop (from 0.966 to 0.228). The bottom trajectory demonstrates how GeoSolver avoids this error to reach a fully grounded conclusion.

to minimize accumulated penalties. Conversely, our Process-Aware (PA) drop-moment mechanism circumvents this length bias by penalizing only sudden relative confidence drops, effectively filtering logical errors while preserving reasoning completeness. Independent of the reward formulation, we investigate the impact of the exploration structure. Upgrading Vanilla GRPO with our entropy-guided tree search yields substantial gains over chain-based rollouts. Under an identical token budget, the tree structure enables a much denser exploration of critical branching points. Ultimately, synergizing these two structural and formulation improvements into the unified Process-Aware Tree-GRPO pipeline achieves the highest overall performance.

Efficacy of PRM Synthesis Strategies. A robust PRM requires both rigorous logical evaluation and precise visual grounding. In Table 6, we ablate the two core synthesis strategies of Geo-PRM-2M: intrinsic MCTS samples (S_{MC}) and Synthetic Hallucination Injection (S_{SHI}). Training the reward model without S_{SHI} diminishes its sensitivity to subtle visual-textual misalignments, confirming that explicit box perturbations are vital for detecting ungrounded coordinates. Similarly, omitting S_{MC} severely limits the model's capacity to evaluate deep, multi-step logical reasoning. The consistent TTS performance drop across both GeoSolver and the generic Qwen3-VL-8B confirms that intrinsic logical mining and synthetic visual perturbations are mutually indispensable.

4.6 Qualitative Analysis

To intuitively illustrate the system-level impact of our approach, Figure 5 demonstrates how process supervision actively intervenes during TTS. While autoregressive generation might blindly commit to a hallucinated object (as seen in the ungrounded Step 5), GeoPRM translates this subtle visual-textual misalignment

into a measurable drop-moment. By integrating these dense, token-level signals, GeoSolver effectively prunes deceptive branches early in the search space, ensuring that the final output is strictly anchored to the visual evidence rather than logical extrapolation.

5 Conclusion

In this paper, we introduce GeoSolver, a comprehensive framework that scales geospatial reasoning through rigorous process supervision. To combat spatial hallucinations, we construct the Geo-PRM-2M dataset by leveraging Entropy-Guided MCTS and Synthetic Hallucination Injection, enabling the training of GeoPRM as a highly calibrated token-level verifier. For policy alignment, we propose Process-Aware Tree-GRPO, an algorithm that synergizes entropy-guided tree exploration with a process-level penalty to effectively mitigate reward hacking and inherent length bias. Consequently, GeoSolver establishes new state-of-the-art benchmarks across diverse remote sensing tasks under standard inference. Furthermore, we demonstrate that GeoPRM unlocks profound Test-Time Scaling capabilities; it not only consistently enhances GeoSolver's performance but also generalizes as a plug-and-play verifier for generalist VLMs. Ultimately, this work establishes a new paradigm for scaling test-time compute in remote sensing, demonstrating that process-guided exploration is essential for achieving verifiable and faithful geospatial intelligence.

Acknowledgements

This work was supported by the National Natural Science Foundation of China under Grant Nos. U22A2098, 62206105 and 62202200; the China Postdoctoral Science Foundation under Grant No. 2025M780312; the Key Science and Technology Development Plan of Jilin Province under Grant No. 20240302078GX. This research was also funded by the Major Science and Technology Development Plan of Changchun under Grant No.2024WX05.

References

1. Anthropic: Claude opus 4 & claude sonnet 4 system card. `https://www.anthropic.com/claude-4-system-card` (2025)
2. Bai, S., Cai, Y., Chen, R., Chen, K., Chen, X., Cheng, Z., Deng, L., Ding, W., Gao, C., Ge, C., Ge, W., Guo, Z., Huang, Q., Huang, J., Huang, F., Hui, B., Jiang, S., Li, Z., Li, M., Li, M., Li, K., Lin, Z., Lin, J., Liu, X., Liu, J., Liu, C., Liu, Y., Liu, D., Liu, S., Lu, D., Luo, R., Lv, C., Men, R., Meng, L., Ren, X., Ren, X., Song, S., Sun, Y., Tang, J., Tu, J., Wan, J., Wang, P., Wang, P., Wang, Q., Wang, Y., Xie, T., Xu, Y., Xu, H., Xu, J., Yang, Z., Yang, M., Yang, J., Yang, A., Yu, B., Zhang, F., Zhang, H., Zhang, X., Zheng, B., Zhong, H., Zhou, J., Zhou, F., Zhou, J., Zhu, Y., Zhu, K.: Qwen3-vl technical report. arXiv preprint arXiv:2511.21631 (2025)

3. Bai, S., Chen, K., Liu, X., Wang, J., Ge, W., Song, S., Dang, K., Wang, P., Wang, S., Tang, J., Zhong, H., Zhu, Y., Yang, M., Li, Z., Wan, J., Wang, P., Ding, W., Fu, Z., Xu, Y., Ye, J., Zhang, X., Xie, T., Cheng, Z., Zhang, H., Yang, Z., Xu, H., Lin, J.: Qwen2.5-vl technical report. arXiv preprint arXiv:2502.13923 (2025)
4. Browne, C.B., Powley, E., Whitehouse, D., Lucas, S.M., Cowling, P.I., Rohlfshagen, P., Tavener, S., Perez, D., Samothrakis, S., Colton, S.: A survey of monte carlo tree search methods. IEEE Transactions on Computational Intelligence and AI in games **4**(1), 1–43 (2012)
5. Cheng, G., Han, J., Lu, X.: Remote sensing image scene classification: Benchmark and state of the art. Proceedings of the IEEE **105**(10), 1865–1883 (2017)
6. Cheng, G., Han, J., Zhou, P., Guo, L.: Multi-class geospatial object detection and geographic image classification based on collection of part detectors. ISPRS Journal of Photogrammetry and Remote Sensing **98**, 119–132 (2014)
7. Cheng, Q., Huang, H., Xu, Y., Zhou, Y., Li, H., Wang, Z.: Nwpu-captions dataset and mlca-net for remote sensing image captioning. IEEE Transactions on Geoscience and Remote Sensing **60**, 1–19 (2022)
8. Cobbe, K., Kosaraju, V., Bavarian, M., Chen, M., Jun, H., Kaiser, L., Plappert, M., Tworek, J., Hilton, J., Nakano, R., et al.: Training verifiers to solve math word problems. arXiv preprint arXiv:2110.14168 (2021)
9. Comanici, G., Bieber, E., Schaekermann, M., Pasupat, I., Sachdeva, N., Dhillon, I., Blistein, M., Ram, O., Zhang, D., Rosen, E., et al.: Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. arXiv preprint arXiv:2507.06261 (2025)
10. DeepSeek-AI: Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning (2025), https://arxiv.org/abs/2501.12948
11. Ding, J., Xue, N., Xia, G.S., Bai, X., Yang, W., Yang, M., Belongie, S., Luo, J., Datcu, M., Pelillo, M., Zhang, L.: Object detection in aerial images: A large-scale benchmark and challenges. IEEE Transactions on Pattern Analysis and Machine Intelligence pp. 1–1 (2021). https://doi.org/10.1109/TPAMI.2021.3117983
12. Fiaz, M., Debary, H., Fraccaro, P., Paudel, D., Van Gool, L., Khan, F., Khan, S.: Geovlm-r1: Reinforcement fine-tuning for improved remote sensing reasoning. arXiv preprint arXiv:2509.25026 (2025)
13. Fini, E., Shukor, M., Li, X., Dufter, P., Klein, M., Haldimann, D., Aitharaju, S., da Costa, V.G.T., Béthune, L., Gan, Z., et al.: Multimodal autoregressive pre-training of large vision encoders. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 9641–9654 (2025)
14. Gao, L., Madaan, A., Zhou, S., Alon, U., Liu, P., Yang, Y., Callan, J., Neubig, G.: Pal: Program-aided language models. In: International conference on machine learning. pp. 10764–10799. PMLR (2023)
15. Guo, X., Lao, J., Dang, B., Zhang, Y., Yu, L., Ru, L., Zhong, L., Huang, Z., Wu, K., Hu, D., He, H., Wang, J., Chen, J., Yang, M., Zhang, Y., Li, Y.: Skysense: A multi-modal remote sensing foundation model towards universal interpretation for earth observation imagery. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 27672–27683 (June 2024)
16. Guo, Y., Xu, L., Liu, J., Ye, D., Qiu, S.: Segment policy optimization: Effective segment-level credit assignment in rl for large language models. arXiv preprint arXiv:2505.23564 (2025)
17. Hou, Z., Hu, Z., Li, Y., Lu, R., Tang, J., Dong, Y.: Treerl: Llm reinforcement learning with on-policy tree search. In: Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 12355–12369 (2025)

18. Hu, H., Wang, P., Feng, Y., Wei, K., Yin, W., Diao, W., Wang, M., Bi, H., Kang, K., Ling, T., et al.: Ringmo-agent: A unified remote sensing foundation model for multi-platform and multi-modal reasoning. arXiv preprint arXiv:2507.20776 (2025)
19. Huang, W., Jia, B., Zhai, Z., Cao, S., Ye, Z., Zhao, F., Xu, Z., Hu, Y., Lin, S.: Vision-r1: Incentivizing reasoning capability in multimodal large language models. arXiv preprint arXiv:2503.06749 (2025)
20. Ji, Y., Ma, Z., Wang, Y., Chen, G., Chu, X., Wu, L.: Tree search for llm agent reinforcement learning. arXiv preprint arXiv:2509.21240 (2025)
21. Jian, D., Nan, X., Yang, L., Gui-Song, X., Lu, Q.: Learning roi transformer for detecting oriented objects in aerial images. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (June 2019)
22. Kuckreja, K., Danish, M.S., Naseer, M., Das, A., Khan, S., Khan, F.S.: Geochat: Grounded large vision-language model for remote sensing. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 27831–27840 (2024)
23. Lenton, T.M., Abrams, J.F., Bartsch, A., Bathiany, S., Boulton, C.A., Buxton, J.E., Conversi, A., Cunliffe, A.M., Hebden, S., Lavergne, T., et al.: Remotely sensing potential climate change tipping points across scales. nature communications **15**(1), 343 (2024)
24. Li, K., Xin, Z., Pang, L., Pang, C., Deng, Y., Yao, J., Xia, G., Meng, D., Wang, Z., Cao, X.: Segearth-r1: Geospatial pixel reasoning via large language model. arXiv preprint arXiv:2504.09644 (2025)
25. Li, X., Ding, J., Elhoseiny, M.: Vrsbench: A versatile vision-language benchmark dataset for remote sensing image understanding. In: Advances in Neural Information Processing Systems. vol. 37, pp. 3229–3242. Curran Associates, Inc. (2024)
26. Li, Z.Z., Zhang, D., Zhang, M.L., Zhang, J., Liu, Z., Yao, Y., Xu, H., Zheng, J., Wang, P.J., Chen, X., et al.: From system 1 to system 2: A survey of reasoning large language models. arXiv preprint arXiv:2502.17419 (2025)
27. Lightman, H., Kosaraju, V., Burda, Y., Edwards, H., Baker, B., Lee, T., Leike, J., Schulman, J., Sutskever, I., Cobbe, K.: Let's verify step by step. In: The twelfth international conference on learning representations (2023)
28. Liu, J., Fu, R., Liu, H., Sun, L., Yang, B.: Geodit: A diffusion-based vision-language model for geospatial understanding. arXiv preprint arXiv:2512.02505 (2025)
29. Liu, J., Fu, R., Sun, L., Liu, H., Yang, X., Zhang, W., Na, X., Duan, Z., Yang, B.: Skymoe: A vision-language foundation model for enhancing geospatial interpretation with mixture of experts. arXiv preprint arXiv:2512.02517 (2025)
30. Liu, J., Sun, L., Fu, R., Yang, B.: Towards faithful reasoning in remote sensing: A perceptually-grounded geospatial chain-of-thought for vision-language models. arXiv preprint arXiv:2509.22221 (2025)
31. Liu, S., Ma, Y., Zhang, X., Wang, H., Ji, J., Sun, X., Ji, R.: Rotated multi-scale interaction network for referring remote sensing image segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 26658–26668 (2024)
32. Lobry, S., Marcos, D., Murray, J., Tuia, D.: Rsvqa: Visual question answering for remote sensing data. IEEE Transactions on Geoscience and Remote Sensing **58**(12), 8555–8566 (2020)
33. Long, Y., Gong, Y., Xiao, Z., Liu, Q.: Accurate object localization in remote sensing images based on convolutional neural networks. IEEE Transactions on Geoscience and Remote Sensing **55**(5), 2486–2498 (2017)

34. Lu, X., Wang, B., Zheng, X., Li, X.: Exploring models and data for remote sensing image caption generation. IEEE Transactions on Geoscience and Remote Sensing **56**(4), 2183–2195 (2017)
35. Luo, J., Pang, Z., Zhang, Y., Wang, T., Wang, L., Dang, B., Lao, J., Wang, J., Chen, J., Tan, Y., Li, Y.: Skysensegpt: A fine-grained instruction tuning dataset and model for remote sensing vision-language understanding (2024), https://arxiv.org/abs/2406.10100
36. Luo, L., Liu, Y., Liu, R., Phatale, S., Guo, M., Lara, H., Li, Y., Shu, L., Zhu, Y., Meng, L., et al.: Improve mathematical reasoning in language models by automated process supervision. arXiv preprint arXiv:2406.06592 (2024)
37. Luo, R., Zheng, Z., Wang, Y., Yu, Y., Ni, X., Lin, Z., Zeng, J., Yang, Y.: Ursa: Understanding and verifying chain-of-thought reasoning in multimodal mathematics. arXiv e-prints pp. arXiv–2501 (2025)
38. Muennighoff, N., Yang, Z., Shi, W., Li, X.L., Fei-Fei, L., Hajishirzi, H., Zettlemoyer, L., Liang, P., Candès, E., Hashimoto, T.B.: s1: Simple test-time scaling. In: Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing. pp. 20286–20332 (2025)
39. OpenAI: Introducing gpt-5. https://openai.com/introducing-gpt-5/ (2025)
40. Pak, H.Y., Lin, W., Law, A.W.K.: Correction of systematic image misalignment in direct georeferencing of uav multispectral imagery. International Journal of Remote Sensing **46**(3), 930–952 (2025)
41. Pang, C., Weng, X., Wu, J., Li, J., Liu, Y., Sun, J., Li, W., Wang, S., Feng, L., Xia, G.S., et al.: Vhm: Versatile and honest vision language model for remote sensing image analysis. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 39, pp. 6381–6388 (2025)
42. Peng, M., Chen, N., Suo, Z., Li, J.: Rewarding graph reasoning process makes llms more generalized reasoners. In: Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 2. pp. 2257–2268 (2025)
43. Peng, S., Fu, D., Gao, L., Zhong, X., Fu, H., Tang, Z.: Multimath: Bridging visual and mathematical reasoning for large language models. arXiv preprint arXiv:2409.00147 (2024)
44. Schulman, J., Wolski, F., Dhariwal, P., Radford, A., Klimov, O.: Proximal policy optimization algorithms. arXiv preprint arXiv:1707.06347 (2017)
45. Shabbir, A., Zumri, M., Bennamoun, M., Khan, F.S., Khan, S.: Geopixel: Pixel grounding large multimodal model in remote sensing. In: International Conference on Machine Learning. pp. 54095–54111. PMLR (2025)
46. Shao, R., Li, Z., Zhang, Z., Xu, L., He, X., Yuan, H., He, B., Dai, Y., Yan, Y., Chen, Y., et al.: Asking like socrates: Socrates helps vlms understand remote sensing images. arXiv preprint arXiv:2511.22396 (2025)
47. Shao, Z., Wang, P., Zhu, Q., Xu, R., Song, J., Bi, X., Zhang, H., Zhang, M., Li, Y., Wu, Y., et al.: Deepseekmath: Pushing the limits of mathematical reasoning in open language models. arXiv preprint arXiv:2402.03300 (2024)
48. Snell, C., Lee, J., Xu, K., Kumar, A.: Scaling llm test-time compute optimally can be more effective than scaling model parameters. arXiv preprint arXiv:2408.03314 (2024)
49. Soni, S., Dudhane, A., Debary, H., Fiaz, M., Munir, M.A., Danish, M.S., Fraccaro, P., Watson, C.D., Klein, L.J., Khan, F.S., et al.: Earthdial: Turning multi-sensory earth observations to interactive dialogues. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 14303–14313 (2025)

50. Sun, Y., Feng, S., Li, X., Ye, Y., Kang, J., Huang, X.: Visual grounding in remote sensing images. In: Proceedings of the 30th ACM International conference on Multimedia. pp. 404–412 (2022)
51. Team, K., Du, A., Yin, B., Xing, B., Qu, B., Wang, B., Chen, C., Zhang, C., Du, C., Wei, C., Wang, C., Zhang, D., Du, D., Wang, D., Yuan, E., Lu, E., Li, F., Sung, F., Wei, G., Lai, G., Zhu, H., Ding, H., Hu, H., Yang, H., Zhang, H., Wu, H., Yao, H., Lu, H., Wang, H., Gao, H., Zheng, H., Li, J., Su, J., Wang, J., Deng, J., Qiu, J., Xie, J., Wang, J., Liu, J., Yan, J., Ouyang, K., Chen, L., Sui, L., Yu, L., Dong, M., Dong, M., Xu, N., Cheng, P., Gu, Q., Zhou, R., Liu, S., Cao, S., Yu, T., Song, T., Bai, T., Song, W., He, W., Huang, W., Xu, W., Yuan, X., Yao, X., Wu, X., Zu, X., Zhou, X., Wang, X., Charles, Y., Zhong, Y., Li, Y., Hu, Y., Chen, Y., Wang, Y., Liu, Y., Miao, Y., Qin, Y., Chen, Y., Bao, Y., Wang, Y., Kang, Y., Liu, Y., Du, Y., Wu, Y., Wang, Y., Yan, Y., Zhou, Z., Li, Z., Jiang, Z., Zhang, Z., Yang, Z., Huang, Z., Huang, Z., Zhao, Z., Chen, Z.: Kimi-VL technical report (2025), https://arxiv.org/abs/2504.07491
52. Team, V., Hong, W., Yu, W., Gu, X., Wang, G., Gan, G., Tang, H., Cheng, J., Qi, J., Ji, J., Pan, L., Duan, S., Wang, W., Wang, Y., Cheng, Y., He, Z., Su, Z., Yang, Z., Pan, Z., Zeng, A., Wang, B., Chen, B., Shi, B., Pang, C., Zhang, C., Yin, D., Yang, F., Chen, G., Xu, J., Zhu, J., Chen, J., Chen, J., Chen, J., Lin, J., Wang, J., Chen, J., Lei, L., Gong, L., Pan, L., Liu, M., Xu, M., Zhang, M., Zheng, Q., Yang, S., Zhong, S., Huang, S., Zhao, S., Xue, S., Tu, S., Meng, S., Zhang, T., Luo, T., Hao, T., Tong, T., Li, W., Jia, W., Liu, X., Zhang, X., Lyu, X., Fan, X., Huang, X., Wang, Y., Xue, Y., Wang, Y., Wang, Y., An, Y., Du, Y., Shi, Y., Huang, Y., Niu, Y., Wang, Y., Yue, Y., Li, Y., Zhang, Y., Wang, Y., Wang, Y., Zhang, Y., Xue, Z., Hou, Z., Du, Z., Wang, Z., Zhang, P., Liu, D., Xu, B., Li, J., Huang, M., Dong, Y., Tang, J.: Glm-4.5v and glm-4.1v-thinking: Towards versatile multimodal reasoning with scalable reinforcement learning (2025), https://arxiv.org/abs/2507.01006
53. Wang, D., Liu, S., Jiang, W., Wang, F., Liu, Y., Qin, X., Luo, Z., Zhou, C., Guo, H., Zhang, J., et al.: Geozero: Incentivizing reasoning from scratch on geospatial scenes. arXiv preprint arXiv:2511.22645 (2025)
54. Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., Le, Q.V., Zhou, D., et al.: Chain-of-thought prompting elicits reasoning in large language models. Advances in neural information processing systems **35**, 24824–24837 (2022)
55. Xia, G.S., Bai, X., Ding, J., Zhu, Z., Belongie, S., Luo, J., Datcu, M., Pelillo, M., Zhang, L.: Dota: A large-scale dataset for object detection in aerial images. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3974–3983 (2018)
56. Xia, G.S., Hu, J., Hu, F., Shi, B., Bai, X., Zhong, Y., Zhang, L., Lu, X.: Aid: A benchmark data set for performance evaluation of aerial scene classification. IEEE Transactions on Geoscience and Remote Sensing **55**(7), 3965–3981 (2017)
57. Xia, G.S., Yang, W., Delon, J., Gousseau, Y., Sun, H., Maître, H.: Structural high-resolution satellite image indexing. In: ISPRS TC VII Symposium-100 Years ISPRS. vol. 38, pp. 298–303 (2010)
58. Xie, B., Xu, B., Yuan, Y., Zhu, S., Shen, H.: From outcomes to processes: Guiding prm learning from orm for inference-time alignment. In: Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 19291–19307 (2025)
59. Yan, Y., Wang, S., Huo, J., Li, H., Li, B., Su, J., Gao, X., Zhang, Y.F., Xu, T., Chu, Z., et al.: Errorradar: Benchmarking complex mathematical reasoning of mul-

timodal large language models via error detection. arXiv preprint arXiv:2410.04509 (2024)
60. Yang, A., Zhang, B., Hui, B., Gao, B., Yu, B., Li, C., Liu, D., Tu, J., Zhou, J., Lin, J., et al.: Qwen2. 5-math technical report: Toward mathematical expert model via self-improvement. arXiv preprint arXiv:2409.12122 (2024)
61. Yang, Y., Newsam, S.: Bag-of-visual-words and spatial extensions for land-use classification. In: ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems (ACM GIS) (2010)
62. Yao, L., Liu, F., Lu, H., Zhang, C., Min, R., Xu, S., Di, S., Peng, P.: Remotereasoner: Towards unifying geospatial reasoning workflow. arXiv preprint arXiv:2507.19280 (2025)
63. Yu, Y., Zhang, Y., Zhang, D., Liang, X., Zhang, H., Zhang, X., Khademi, M., Awadalla, H.H., Wang, J., Yang, Y., et al.: Chain-of-reasoning: Towards unified mathematical reasoning in large language models via a multi-paradigm perspective. In: Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 24914–24937 (2025)
64. Yuan, Z., Zhang, W., Fu, K., Li, X., Deng, C., Wang, H., Sun, X.: Exploring a fine-grained multiscale method for cross-modal remote sensing image retrieval. IEEE Transactions on Geoscience and Remote Sensing **60**, 1–19 (2021)
65. Zhan, Y., Xiong, Z., Yuan, Y.: Rsvg: Exploring data and models for visual grounding on remote sensing data. IEEE Transactions on Geoscience and Remote Sensing **61**, 1–13 (2023)
66. Zhang, L., Hosseini, A., Bansal, H., Kazemi, M., Kumar, A., Agarwal, R.: Generative verifiers: Reward modeling as next-token prediction. arXiv preprint arXiv:2408.15240 (2024)
67. Zhang, W., Cai, M., Zhang, T., Zhuang, Y., Li, J., Mao, X.: Earthmarker: A visual prompting multi-modal large language model for remote sensing. IEEE Transactions on Geoscience and Remote Sensing (2024)
68. Zhang, W., Cai, M., Zhang, T., Zhuang, Y., Mao, X.: Earthgpt: A universal multi-modal large language model for multi-sensor image comprehension in remote sensing domain. IEEE Transactions on Geoscience and Remote Sensing (2024)
69. Zhang, Y., Yuan, Y., Feng, Y., Lu, X.: Hierarchical and robust convolutional neural network for very high-resolution remote sensing object detection. IEEE Transactions on Geoscience and Remote Sensing **57**(8), 5535–5548 (2019)
70. Zhang, Z., Guan, Z., Zhao, T., Shen, H., Li, T., Cai, Y., Su, Z., Liu, Z., Yin, J., Li, X.: Geo-r1: Improving few-shot geospatial referring expression understanding with reinforcement fine-tuning. arXiv preprint arXiv:2509.21976 (2025)
71. Zhao, B., Zhong, Y., Xia, G., Zhang, L.: Dirichlet-derived multiple topic scene classification model fusing heterogeneous features for high spatial resolution remote sensing imagery. IEEE Trans. Geosci. Remote Sens **54**(4), 2108–2123 (2016)
72. Zhao, B., Zhong, Y., Zhang, L., Huang, B.: The fisher kernel coding framework for high spatial resolution scene classification. Remote Sensing **8**(2), 157 (2016)
73. Zheng, C., Zhu, J., Ou, Z., Chen, Y., Zhang, K., Shan, R., Zheng, Z., Yang, M., Lin, J., Yu, Y., et al.: A survey of process reward models: From outcome signals to process supervisions for large language models. arXiv preprint arXiv:2510.08049 (2025)
74. Zhu, Q., Lao, J., Ji, D., Luo, J., Wu, K., Zhang, Y., Ru, L., Wang, J., Chen, J., Yang, M., Liu, D., Zhao, F.: Skysense-o: Towards open-world remote sensing interpretation with vision-centric visual-language modeling. In: Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR). pp. 14733–14744 (June 2025)

75. Zhu, Q., Zhong, Y., Zhao, B., Xia, G.S., Zhang, L.: Bag-of-visual-words scene classifier with local and global features for high spatial resolution remote sensing imagery. IEEE Geoscience and Remote Sensing Letters **13**(6), 747–751 (2016)
76. Zhuang, W., Huang, X., Zhang, X., Zeng, J.: Math-puma: Progressive upward multimodal alignment to enhance mathematical reasoning. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 39, pp. 26183–26191 (2025)