TERRAEVO: SKILL-GUIDED SELF-EVOLUTION OF EARTH OBSERVATION VISION-LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Remote-sensing vision-language models (VLMs) remain heavily dependent on human-annotated supervision, constraining their scalability across increasingly diverse Earth observation (EO) imagery. Self-evolution offers a promising alternative, yet sustaining the reliability and diversity of self-generated supervision remains challenging. We introduce **TerraEvo**, a self-evolving framework that enables Proposer and Solver roles to co-evolve from unlabeled EO imagery without additional human annotations or an external judge. At its core, skill-guided Socratic dialogue transforms the Solver's answer disagreement into targeted follow-up questions. Quality supervision and adaptive rewards encourage informative, visually grounded questioning, while skill balancing and knowledge preservation support sustained learning across diverse EO tasks. A composite label-free learning signal further integrates answer consistency and reasoning quality to guide Solver improvement. Across five remote-sensing benchmarks, TerraEvo achieves the highest average accuracy among the evaluated general-purpose and domain-specific models, while ablation studies demonstrate the effectiveness of its key components. Our code, data, and models will be released publicly.

1 INTRODUCTION

Interpreting complex scenes from high-resolution satellite and aerial imagery is a key challenge in remote sensing, with important applications in change detection, land-use classification, urban planning, environmental monitoring, and disaster response (Fiaz et al., 2025b; Danish et al., 2026; Soni et al., 2025; Gupta et al., 2019b). As the volume of EO imagery grows, models must understand how objects and their surroundings relate to one another. For example, answering *"Which buildings are cut off from the main road by floodwater?"* requires identifying buildings, tracing their access roads, and determining where floodwater interrupts these connections. Recognizing objects alone is insufficient to answer such questions, which depend on spatial relationships and scene conditions. Recent EO vision-language models (VLMs) have made progress in multimodal understanding and reasoning (Kuckreja et al., 2024; Fiaz et al., 2025a; Liu et al., 2026b; Wang et al., 2026a), with improved performance on visual reasoning benchmarks (Wang et al., 2026b; Yao et al., 2026). However, these advances rely heavily on human-labeled training data, limiting their scalability across diverse EO tasks and scenes. Although EO imagery is increasingly available, high-quality annotated data remain scarce, especially data that jointly connect images, questions, and answers for reasoning-intensive tasks. This gap motivates our central research question. *"Can VLMs progressively improve their capabilities using self-generated learning signals from unlabeled Earth observation (EO) imagery?"*

Self-evolution in Large Language Models (LLMs) has emerged as a popular mechanism that enables models to generate training signals through self-play (Liu et al., 2025) and self-critique (Wu et al., 2025). Several approaches organize learning as a closed loop between the proposal and the solver roles (Huang et al., 2026; Zhao et al., 2026; Yang et al., 2026), allowing self-driven improvement without additional human annotation. Recent work extends this approach to VLMs, enabling them to advance through self-generated learning signals, such as (He et al., 2026; Li et al., 2026b). However, reliance on self-consistency can reinforce incorrect predictions when agreement is treated as evidence of correctness (Thawakar et al., 2025; He et al., 2026; Sunil et al., 2026).

Generating reliable supervision from images introduces a fundamental vulnerability, as errors in visual understanding can affect both question generation and answering. Without appropriate qual-

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.

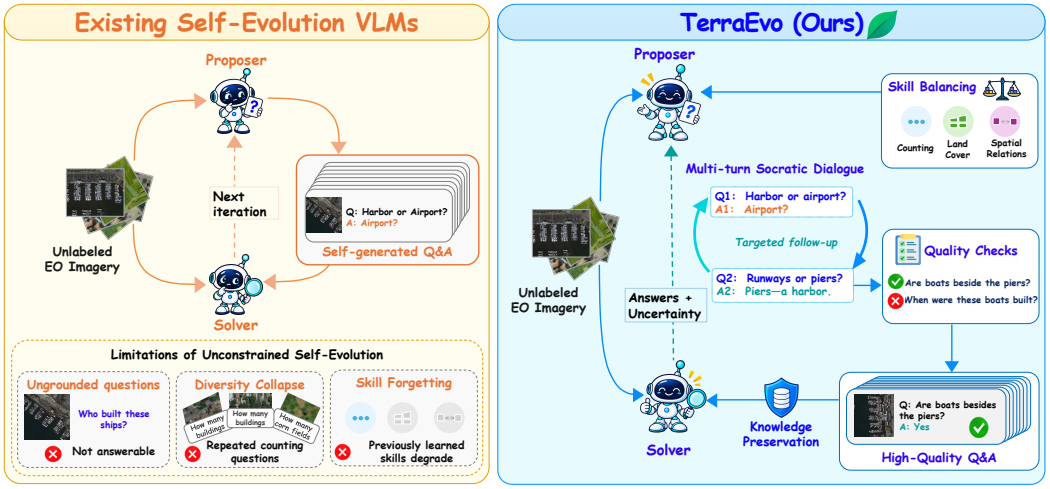


Figure 1: **Comparison of existing self-evolving VLMs and TerraEvo.** Unconstrained self-evolution risks ungrounded questions, limited skill diversity, and skill forgetting. TerraEvo combines Socratic dialogue, skill balancing, quality supervision, and knowledge preservation to support reliable self-evolution from unlabeled EO imagery.

ity and diversity controls, self-evolving training may produce repetitive, poorly grounded questions and gradually concentrate on a narrow set of easy skills. These vulnerabilities are particularly challenging in EO imagery, where scenes contain densely packed small objects, complex spatial relationships, seasonal and illumination changes, and limited coverage of rare events in the scenes. As a result, unconstrained self-evolution may favor shallow, low-diversity questions that fail to capture the range of skills required for EO understanding, as illustrated in Fig. 1. Although answer agreement provides a useful learning signal, it does not by itself ensure visual grounding, sound reasoning, or balanced skill coverage (Thawakar et al., 2025; He et al., 2026). These limitations motivate a Proposer–Solver framework that strengthens the reliability of self-generated supervision while maintaining question quality and skill diversity.

To address these challenges, we introduce TerraEvo, a self-evolving framework that co-evolves Proposer and Solver roles from unlabeled EO imagery without additional human annotations or an external judge. The proposed framework enables their interaction such that each interaction targets an under-represented skill and follows a multi-turn Socratic dialogue, where the Proposer conditions subsequent questions on disagreement among the Solver's sampled answers. An adaptive question reward calibrates question difficulty using the Solver's answer agreement while accounting for skill coverage, validity, and repetition. The highest-reward question across the dialogue is selected for training. TerraEvo combines quality supervision and skill balancing to improve reliability and diversity of self-generated questions. A sliding-window skill tracker guides question generation through an under-representation bonus and targeted regeneration. A content-based router checks skill assignments, while the Solver assesses question grounding, answerability, and validity. The Solver learns through a composite reward that combines soft answer agreement, trajectory-level agreement, and CoT quality. Skill-adaptive KL regularization limits policy drift, while skill-balanced replay maintains exposure to previously encountered skills. Per-skill confidence thresholds and validity checks govern admission to the replay buffer. Our main contributions are summarized below.

- We introduce TerraEvo, a self-evolving framework for EO VLMs that combines skill-targeted Socratic dialogue with adaptive question rewards. Follow-ups use the Solver's answer disagreement to refine questions, and training selects the highest-reward question across the dialogue.
- We develop quality and diversity controls that combine content-based skill verification, question-validity checks, and skill-aware generation to mitigate poorly grounded questions and narrow skill coverage.

- We design a label-free composite Solver reward that evaluates answer agreement, trajectory agreement, and CoT quality, complemented by skill-adaptive KL regularization and skill-balanced replay for knowledge preservation.

2 Related Work

2.1 Earth Observation Vision-Language Models

Advances in general-purpose vision-language models (VLMs) (Bai et al., 2025b;a; Zhu et al., 2025; Chen et al., 2024b) have catalyzed their rapid adaptation to satellite and aerial imagery, enabling natural-language interaction with Earth observation (EO) data. GeoChat (Kuckreja et al., 2024) and EarthGPT (Zhang et al., 2024) support geospatial dialogue capabilities across diverse tasks, extending interaction with overhead imagery beyond fixed, task-specific outputs. Building on this foundation, VHM (Pang et al., 2025), GeoLLaVA-8K (Wang et al., 2026a), EarthDial (Soni et al., 2025), TEOChat (Irvin et al., 2025), SkySenseGPT (Luo et al., 2024), and LHRS-Bot (Muhtar et al., 2024) further expanded the capabilities and task coverage of EO VLMs. More recently, Geo-R1 (Zhang et al., 2026), GeoVLM-R1 (Fiaz et al., 2025a), and RSThinker (Liu et al., 2026a) have explored reinforcement fine-tuning and explicit reasoning to improve performance on EO tasks. Despite these advances, constructing suitable training data and supervision remains a major challenge as EO imagery grows in volume and geographic coverage.

2.2 Self-Evolving Vision-Language Models

Self-evolution enables large language models to improve through self-generated training data and feedback, reducing reliance on additional human annotation (Liu et al., 2025; Zhao et al., 2026; Wu et al., 2025; Yang et al., 2026). Several VLM frameworks extend this idea through interacting question-generation and answer-generation roles. EvoLMM (Thawakar et al., 2025) jointly evolves question generation and answering using continuous self-consistency rewards. VisPlay (He et al., 2025) combines uncertainty-based question rewards with diversity regularization in an image-conditioned Questioner–Reasoner framework. C2-Evo (Chen et al., 2025) jointly evolves multimodal training data and model capabilities through adaptive data generation and selection. DoGe (Li et al., 2026a) separates context understanding from problem solving through Thinker and Solver components. Vision-Zero (Wang et al., 2026c) introduces dependencies on external tools, such as the image-editing modules used. DUEL (Qiu et al., 2026) uses adversarial self-play between a Challenger that generates paired claims and a Solver that verifies them against the image.

Limitations. These approaches address complementary aspects of self-evolution, but answer agreement alone cannot establish visual correctness or ensure skill coverage. In EO imagery, self-generated questions may be poorly grounded or favor familiar tasks, leaving other skills underrepresented, and continued training on such skewed data can erode prior knowledge. TerraEvo addresses these challenges through skill-targeted, Solver-conditioned multi-turn dialogue, content-based skill verification, question-validity checks, and composite label-free rewards, while skill-adaptive KL regularization and skill-balanced replay preserve knowledge across skills.

3 Preliminaries

Let a Proposer π_q generate a question q for an image input x, and a Solver π_s produce a response o given context $c = (x, q)$. Each role learns from its own scalar reward and is regularized toward a reference policy π_{ref}.

Solver learning. The Solver π_s uses a Group Relative Policy Optimization (GRPO)-based update that compares rewards across responses to the same context. For a group of $G > 1$ responses $\{o_j\}_{j=1}^{G}$ with rewards $\{r_j\}_{j=1}^{G}$, the normalized advantage is

$$\hat{A}_j = \frac{r_j - \mu_r}{\sigma_r + \varepsilon}, \tag{1}$$

where μ_r and σ_r denote the group reward mean and standard deviation, and $\varepsilon = 10^{-8}$ ensures numerical stability. The advantage-weighted objective combines learning with reference-policy reg-

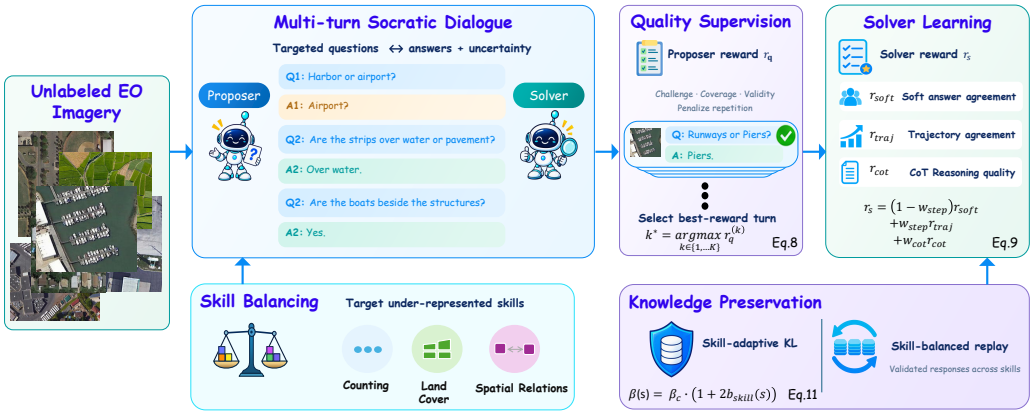


Figure 2: **Overview of TerraEvo.** A skill-guided Proposer generates questions through Solver-conditioned Socratic dialogue on unlabeled EO imagery. The Proposer reward r_q favors challenging, diverse, and valid questions, and the turn with the highest reward is selected for training and periodic Proposer updates (Eq. 8). The Solver is trained with a composite reward r_s that combines answer agreement, trajectory agreement, and reasoning quality (Eq. 9). Skill-adaptive KL $\beta(s)$ (Eq. 11) and skill-balanced replay support knowledge preservation.

ularization.

$$\mathcal{L}_{\text{Solver}}(\theta_s) = -\frac{1}{G}\sum_{j=1}^{G} \hat{A}_j \, \overline{\log \pi_{\theta_s}(o_j \mid c)} + \beta_s \, \overline{D_{\text{KL}}}(\pi_{\theta_s} \| \pi_{\text{ref}}), \tag{2}$$

where the overline denotes token averaging to normalize for response length, and β_s balances reward-driven learning with regularization toward the reference policy to limit policy drift.

Proposer learning. The Proposer π_q is optimized using REINFORCE with reward r_q for a generated output o_q. A moving-average baseline b centers the reward to obtain the advantage $r_q - b$. The objective combines reward-weighted learning with reference-policy regularization:

$$\mathcal{L}_{\text{Proposer}}(\theta_q) = (r_q - b) \cdot \text{CE}(\theta_q) + \beta_p \, D_{\text{KL}}(\pi_{\theta_q} \| \pi_{\text{ref}}), \qquad b \leftarrow m_b b + (1 - m_b) r_q, \tag{3}$$

where $\text{CE}(\theta_q)$ is the token-averaged negative log-likelihood of the generated output given its input context. The coefficient m_b controls baseline smoothing, while β_p balances reward-driven learning with regularization toward the reference policy.

4 Method

We introduce TerraEvo, a self-evolving framework that improves EO vision-language models through self-generated supervision from unlabeled imagery, without additional human annotations or an external judge. As illustrated in Fig. 2, a *Proposer* (π_q) and a *Solver* (π_s) share a VLM backbone with separate LoRA adapters. Skill balancing directs the Proposer toward under-represented skills, while multi-turn Socratic dialogue conditions follow-up questions on the Solver's answer disagreement. Quality supervision assesses question validity, and an adaptive Proposer reward guides selection of the highest-reward question and Proposer learning. The Solver learns from its sampled responses through a composite reward that combines answer agreement, trajectory agreement, and chain-of-thought (CoT) reasoning quality. Skill-adaptive regularization and balanced replay complement these updates while helping preserve knowledge across skills.

4.1 Skill Balancing

Unconstrained question generation may favor familiar tasks, leaving other skills under-represented. TerraEvo promotes coverage across eight skill categories $\mathcal{S}$ comprising counting, classification, spatial, assessment, attributes, temporal, reasoning, and scene. A skill tracker monitors recent question counts for each skill and guides the Proposer toward under-represented categories.

Skill bonus. The tracker maintains a sliding window of the W most recently resolved skill labels, including those of questions that fail the validity check. Let n_s denote the count for skill s and

$\bar{n} = \frac{1}{|\mathcal{S}|} \sum_{s'} n_{s'}$ the mean count across skills. The skill bonus is

$$b_{\text{skill}}(s) = \max\left(\frac{\bar{n} - n_s}{\bar{n}}, 0\right), \tag{4}$$

which assigns larger bonuses to skills with lower representation. The bonus is initialized to one for every skill when the tracking window is empty.

Targeted generation and hard-forcing. For each dialogue, we select $s^{\text{tgt}} = \arg\max_s b_{\text{skill}}(s)$ as the target skill. When its bonus exceeds $\tau_{\text{force}} = 0.40$, we activate hard-forcing through a target-skill instruction and up to $R_{\text{max}} = 3$ generation attempts per turn. The first question whose resolved label matches the target is retained, or the final attempt if none matches. After at least 50 observations, a diversity hint also discourages skills with counts above $1.6\bar{n}$, except for the current target.

4.2 Multi-Turn Socratic Interaction

A single question offers limited scope for addressing the Solver's uncertainty. Given the target skill s^{tgt} selected in Section 4.1, TerraEvo runs up to K_{max} dialogue turns for each image record x. At turn k, the Solver samples M responses to question q_k. Their normalized answers define the empirical distribution $p(a \mid x, q_k)$, where each probability is the fraction of samples producing answer a. The majority answer is $\hat{a}_k = \arg\max_a p(a \mid x, q_k)$, and its agreement score is $u_k = u(x, q_k) = \max_a p(a \mid x, q_k)$. The dialogue ends when $u_k \geq \tau_{\text{stop}}$ or after K_{max} turns. Otherwise, the Proposer receives a feedback summary f_k highlighting disagreement between the majority answer and minority answers from turn k. This feedback guides a targeted follow-up that asks the Solver to distinguish between the competing answers:

$$q_{k+1} \sim \pi_q(\cdot \mid x, s^{\text{tgt}}, f_k), \qquad k < K_{\text{max}}. \tag{5}$$

The target skill and the regeneration decision established in Section 4.1 remain fixed throughout the dialogue. Answer agreement serves as a proxy for the Solver's uncertainty but does not guarantee correctness.

4.3 Quality Supervision and Proposer Reward

Reliable self-generated supervision requires visually grounded questions and accurate skill assignments. Complementing skill balancing (Section 4.1), a quality supervisor assesses question validity and skill consistency. These assessments contribute to the Proposer reward used to select the best question across dialogue turns (Section 4.3.2).

4.3.1 Quality Supervisor

The quality supervisor checks both the skill represented by a question and whether it supports useful visual supervision. Since the skill assigned to a question determines its skill bonus, accepting the Proposer's declared tag without verification could reward incorrect assignments. A keyword-based router therefore checks the tag against the question's content. For example, "How many buildings are surrounded by floodwater?" combines a disaster-related context with a counting task, which may lead the Proposer to assign an *assessment* tag. The router prioritizes the cue "how many" and resolves the primary skill as *counting*, because the requested answer is a quantity rather than a judgment of flood severity. If no category matches, the router retains a declared tag from the supported taxonomy or defaults to *scene*. The resolved label guides skill tracking, skill-bonus computation, and replay assignment, reducing the risk of counting a question toward the wrong skill.

Skill verification is followed by question validation to assess whether the proposed task is supported by the image. Given the image, question, and resolved skill $\hat{s}$, the Solver jointly evaluates visual grounding, semantic validity, answerability from the image alone, and consistency with the assigned skill. This assessment produces a binary signal $\rho(x, q, \hat{s}) \in \{0, 1\}$ that contributes to the Proposer reward. For questions that pass validation and the skill-specific confidence gate (Section 4.4.2), a separate Solver query checks whether the majority answer is supported by the image. The answer-verification check governs replay admission but does not contribute to the Proposer reward. Separating these checks is important because a valid question may still receive an incorrect majority answer. During the first $S_{\text{warm}} = 200$ steps, both checks are bypassed to support initial training.

After warmup, replay admission requires both checks to pass and the confidence gate to be satisfied. Failure of question validation removes the validity bonus but does not automatically exclude the question from the regular training update. Together, these internal assessments provide quality signals without additional human annotations or an external judge.

4.3.2 PROPOSER REWARD

The question reward encourages appropriate difficulty, broader skill coverage, and validity:

$$r_q(x, q, \hat{s}) = \min\Big(1,\ r_{\text{diff}}(x, q) + \lambda\, b_{\text{skill}}(\hat{s}) + \lambda\, \rho(x, q, \hat{s})\Big), \tag{6}$$

where $\rho(x, q, \hat{s})$ is the validity signal and λ weights both the skill and validity bonuses. The difficulty term calibrates question difficulty using the Solver's recent answer agreement. It rewards agreement below an adaptive reference, favors answer entropy near a target level, and penalizes repeated questions:

$$r_{\text{diff}}(x, q) = \text{clip}\Big(\alpha_d\, \text{clip}\big(\tfrac{\tau_{\text{ref}} - u(x,q)}{\tau_{\text{ref}}}, 0, 1\big) + \alpha_e\, e^{-\frac{(H(x,q)-\mu_H)^2}{2\sigma_H^2}} - \alpha_b\, P_{\text{BLEU}}(q),\ 0,\ 1\Big), \tag{7}$$

where $\tau_{\text{ref}} = \min(1, \bar{e} + \delta)$ is the adaptive agreement reference, with δ denoting the agreement margin. Here, $\bar{e}$ is an exponential moving average of answer agreement with smoothing coefficient m_e, updated as $\bar{e} \leftarrow m_e \bar{e} + (1 - m_e) u(x, q)$ before scoring. The entropy $H(x, q)$ measures variation in sampled answers, and $P_{\text{BLEU}}(q)$ discourages repetition within the dialogue. The difficulty reward is bounded in $[0, 1]$ and contributes to the Proposer reward used for question selection and policy updates. For each of the K completed turns, where $K \le K_{\max}$, let $\hat{s}_k$ be the resolved skill of question q_k, and $r_q^{(k)} = r_q(x, q_k, \hat{s}_k)$ be its Proposer reward. We then select the question with the highest reward

$$k^\star = \underset{k \in \{1,\dots,K\}}{\arg\max}\ r_q^{(k)}, \qquad q^\star = q_{k^\star}. \tag{8}$$

The selected question may come from any completed turn, and its existing Solver responses are reused for training.

4.4 SOLVER LEARNING AND KNOWLEDGE PRESERVATION

4.4.1 COMPOSITE SOLVER REWARD

For the selected question $q^\star$, we evaluate each sampled response using a composite label-free reward that combines soft answer agreement (r_{soft}), trajectory agreement (r_{traj}), and chain-of-thought quality (r_{cot}):

$$r_s^{(j)} = (1 - w_{\text{step}}(t))\, r_{\text{soft}}^{(j)} + w_{\text{step}}(t)\, r_{\text{traj}}^{(j)} + w_{\text{cot}}(t)\, r_{\text{cot}}^{(j)}, \tag{9}$$

where $w_{\text{step}}(t)$ and $w_{\text{cot}}(t)$ gradually increase the contribution of trajectory agreement and CoT quality after warmup. Responses are ranked by their composite rewards, and the highest-reward half are retained for Solver learning. The resulting group of $G = M - \lfloor M/2 \rfloor$ responses receives one gradient update using the advantages defined in Eq. 1 and the objective in Eq. 2.

Soft answer agreement. We assign each response a continuous reward based on how frequently its answer appears among the M samples: $r_{\text{soft}}^{(j)} = p(a_j \mid x, q^\star)^\gamma \big(1 - w_\ell\, \text{clip}(\frac{\ell_j - L_{\text{tgt}}}{L_{\text{tgt}}}, 0, 1)\big)$, where a_j is the answer produced by response j and $p(a_j \mid x, q^\star)$ is the fraction of sampled responses producing answer a_j, while γ controls the emphasis on agreement. The length penalty, weighted by w_ℓ, discourages reasoning beyond the target word count L_{tgt}, with ℓ_j denoting the reasoning length of response j. Unlike a binary majority-match reward, this formulation also credits minority answers according to their frequency.

Trajectory agreement. We compare reasoning embeddings across majority-answer responses:

$$r_{\text{traj}}^{(j)} = \begin{cases} d^{\gamma_d} s_j, & j \in \mathcal{D}_v,\ |\mathcal{D}_v| \ge N_{\min}, \\ 0, & \text{otherwise}, \end{cases} \tag{10}$$

where $\mathcal{D}$ is the majority-answer group, $\mathcal{D}_v$ contains its responses with nonempty reasoning, $N_{\min}$ is the minimum number of such responses, and $d = |\mathcal{D}|/M$. The score $s_j \in [0, 1]$ measures

transformed cosine similarity to the normalized mean reasoning embedding of $\mathcal{D}_v$, while γ_d controls density weighting. This rewards reasoning consistency without verifying correctness.

Chain-of-thought quality. The reward $r_{\text{cot}}^{(j)} \in [0, 0.40]$ encourages explicit and specific reasoning through four components: *presence*, *length*, *consistency*, and *specificity*. A response is eligible only if it contains a `Think:` block of at least 15 words. Presence contributes 0.15, while length contributes up to 0.10, increasing with the length of the reasoning and reaching its maximum at 150 words. Consistency contributes 0.10 when the stated answer matches the majority answer and is mentioned in the reasoning. Specificity contributes 0.05 when at least two types of visual cues are detected, including numbers, percentages, colors, sizes, or conditions. These signals reward reasoning quality through textual cues without directly verifying the cited visual evidence.

4.4.2 Knowledge Preservation

Skill-adaptive KL penalty. To limit policy drift when learning from under-represented skills, we scale the Solver's reference regularization using the same skill bonus

$$\beta(s) = \beta_c \cdot (1 + 2\, b_{\text{skill}}(s)), \tag{11}$$

where $\beta(s)$ replaces the Solver coefficient β_s in Eq. 2, with β_c denoting the adaptive base coefficient. Under-represented skills receive stronger regularization toward the reference policy, supporting knowledge preservation during learning. This scaling applies only to the Solver.

Skill-balanced replay. To maintain exposure to skills that appear less often during training, we store up to 50 retained responses per skill. After the supervision warmup, responses enter the buffer only when their question and majority answer pass the quality checks, and their answer agreement falls within a skill-specific range. The lower bound increases for skills with low agreement over their latest 50 dialogue turns, while the upper bound excludes highly consensual questions. A default agreement estimate is used when history is limited. Every F training steps, we sample one response from each non-empty skill buffer, skipping entries whose question matches the current training question. Each replay update uses one stored response, with its stored reward r_{replay} serving directly as the advantage, together with the current skill-adaptive regularization. These updates help preserve knowledge across skills.

4.5 Overall Training Procedure

At each training step, TerraEvo samples an image record, selects an under-represented target skill, and runs a multi-turn Socratic dialogue. Each turn combines skill verification, Solver answer sampling, quality assessment, and question reward computation. After the dialogue, the highest-reward question is selected, and its existing responses are evaluated using the composite Solver reward. The highest-reward half form the regular Solver update, while eligible responses are also stored in the corresponding skill buffer. Every F steps, TerraEvo performs skill-balanced replay updates for the Solver, followed by a Proposer update using the selected question's generated output and reward (Eq. 3). The two roles adapt their regularization coefficients independently. The complete training procedure is provided in Algorithm 1 in Appendix A.3.

5 Experimental Results

5.1 Implementation Details

We initialize both the Proposer and Solver with Qwen3-VL-4B-Instruct (Bai et al., 2025a), using separate LoRA adapters on a shared backbone (rank 16, $\alpha_{\text{LoRA}} = 32$, dropout 0.05, applied to all attention and MLP projections), with the vision encoder frozen. Training runs for 3K steps on 4 NVIDIA A100 80GB GPUs in bfloat16 precision with AdamW, a learning rate of 10^{-6}, weight decay of 0.01, and gradient clipping at 1.0. The Solver generates $M = 14$ responses per question. The Socratic chain runs for up to $K_{\text{max}} = 5$ turns with an answer-agreement stopping threshold of $\tau_{\text{stop}} = 0.70$ (Section 4.2), and the Proposer is updated every $F = 2$ training steps. Further details are provided in Appendix A.1.

Our training dataset consists of approximately 6,000 unlabeled images, with equal sampling across ten remote-sensing datasets and random sampling within each dataset: NWPU-RESISC (Cheng

Table 1: Zero-shot accuracy (%) across five remote-sensing (RS) VQA benchmarks, comparing TerraEvo with closed-source, open-source, and domain-specific RS MLLMs. Average denotes the unweighted mean across benchmarks. Bold indicates the best result in each column, including ties.

Model	GeoBench-VLM	RSHR-Bench	XLRS-Bench	MME-RealWorld-RS	LHRS-Bench	Average
Closed-Source MLLMs						
GPT-4o (Hurst et al., 2024)	41.1	53.9	44.5	49.7	66.9	51.2
GPT-5 (Singh et al., 2025)	42.7	54.5	43.2	54.2	**74.2**	53.8
Open-Source MLLMs						
InternVL2.5-8B (Chen et al., 2024a)	37.7	54.2	46.4	47.2	61.4	49.4
InternVL3.5-8B (Wang et al., 2025b)	37.9	52.2	51.0	47.5	59.2	49.6
InternVL3-8B (Zhu et al., 2025)	40.4	49.1	45.0	50.9	61.9	49.5
LLaVA1.5-7B (Liu et al., 2024)	29.5	28.4	26.5	22.2	42.1	29.7
LLaVA-OneVision-7B (Li et al., 2024)	36.5	48.0	43.9	49.2	63.0	48.1
Qwen2.5-VL-3B (Bai et al., 2025b)	36.4	47.4	42.1	30.5	61.3	43.5
Qwen2.5-VL-7B (Bai et al., 2025b)	41.7	49.7	45.8	39.7	58.9	47.2
Qwen3-VL-4B (Bai et al., 2025a)	39.3	54.2	49.6	51.2	64.1	51.7
Qwen3-VL-8B (Bai et al., 2025a)	42.2	53.8	50.1	**57.1**	62.4	53.1
Domain-Specific Remote Sensing MLLMs						
GeoChat (Kuckreja et al., 2024)	19.5	29.6	26.6	27.9	31.1	26.9
EarthDial (Soni et al., 2025)	32.3	45.8	39.7	26.6	51.3	39.1
VHM (Pang et al., 2025)	33.7	29.5	26.4	22.9	45.1	31.5
GeoLLaVA-8K (Wang et al., 2026a)	29.7	46.1	51.4	28.0	36.8	38.4
ZoomEarth (Liu et al., 2026b)	31.4	42.4	30.9	38.2	47.4	38.1
Geoeyes (Wang et al., 2026b)	35.6	**55.0**	**54.2**	43.5	61.8	50.0
GeoVista (Zhu et al., 2026)	38.4	51.9	50.0	49.8	55.4	49.1
Self-Evolving						
TerraEvo (Ours)	**43.6**	**55.0**	50.4	55.9	66.1	**54.2**

et al., 2017), PatternNet (Zhou et al., 2018), FMoW (Christie et al., 2018), FGSCR-42 (Di et al., 2021), DOTA (Ding et al., 2021), DIOR (Li et al., 2020), FAIR-1M (Sun et al., 2022), DIOR-RSVG (Zhan et al., 2023), RSVQA (Lobry et al., 2020), and xBD (Gupta et al., 2019a). These datasets span scene classification, object detection, ship and aircraft recognition, visual grounding, visual question answering, and building-damage assessment, providing varied visual contexts for skill-targeted question generation. TerraEvo generates its training questions and learning signals from these images without using the datasets' provided metadata or label annotations.

5.2 Experimental Results

Table 1 compares TerraEvo with closed-source, general-purpose open-source, and domain-specific remote-sensing MLLMs on GeoBench-VLM (Danish et al., 2025), RSHR-Bench (Dang et al., 2025), XLRS-Bench (Wang et al., 2025a), MME-RealWorld-RS (Zhang et al., 2025), and LHRS-Bench (Muhtar et al., 2024). TerraEvo achieves the highest average accuracy of 54.2% across five benchmarks, outperforming all evaluated closed-source, open-source, and remote-sensing-specific MLLMs. GPT-5 leads the closed-source models at 53.8%, while Qwen3-VL-8B and Geoeyes lead the open-source and remote-sensing-specific groups, obtaining scores of 53.1% and 50.0%, respectively. Compared with its Qwen3-VL-4B backbone, TerraEvo increases average accuracy from 51.7% to 54.2%, a gain of 2.5 percentage points, without additional human annotations or an external judge during self-evolution.

Table 2: Component ablations of TerraEvo, reporting average accuracy (%). All baselines use the same Qwen3-VL-4B-Instruct backbone.

Configuration	Acc. (%)
Zero-shot backbone	39.3
Naive self-play (baseline)	40.2
EvoLMM	40.1
iReasoner	40.3
TerraEvo (Ours)	**43.6**
Optimization	
Frozen Proposer	42.6
Retain top 25% of Solver responses	42.8
Retain all Solver responses	43.1
Quality Supervision (§4.3.1)	
w/o question-validity check ρ	41.8
Skill Balancing (§4.1)	
w/o skill guidance (generic Proposer)	41.8
w/o hard-forcing	42.9
w/o diversity hint	42.3
w/o skill bonus	42.8
Knowledge Preservation (§4.4.2)	
w/o skill-adaptive regularization	42.1
Uniform replay instead of skill-balanced replay	42.9
Solver Reward (§4.4.1)	
w/o trajectory agreement r_{traj}	41.9
w/o CoT reward r_{cot}	42.0
Combined Ablation	
Frozen Proposer and w/o CoT reward	40.7

5.3 Ablation Studies

We evaluate the contribution of each TerraEvo component by removing individual mechanisms while keeping the backbone VLM, training data, and training budget of 3K steps fixed. Table 2 compares TerraEvo with baselines and ablated configurations, all using the same Qwen3-VL-4B-Instruct backbone. All results report average accuracy (%) on GeoBench-VLM.

Overall contributions. The zero-shot backbone achieves 39.3%, and single-round naive self-play with answer-agreement supervision improves this only slightly to 40.2%. Prior self-evolving methods yield similarly small gains. EvoLMM, which uses continuous self-consistency rewards, reaches 40.1%, and iReasoner, which adds trajectory agreement, reaches 40.3%. TerraEvo achieves 43.6%, the highest accuracy among all methods and a clear margin over every baseline.

Optimization and quality supervision. Freezing the Proposer reduces accuracy to 42.6%, supporting the benefit of adapting question generation alongside Solver learning. Retaining the highest-reward 25% of Solver responses yields 42.8%, while retaining all responses yields 43.1%. TerraEvo achieves 43.6% by retaining the top half of rollouts. Removing the question-validity check $\rho(x, q, \hat{k})$ reduces accuracy to 41.8%, indicating that validity supervision provides useful guidance beyond answer agreement.

Skill balancing and knowledge preservation. Replacing skill-guided generation with a generic Proposer removes the skill-aware prompt, diversity hint, hard-forcing, and skill-bonus reward, reducing accuracy to 41.8%. Individually removing hard-forcing, the diversity hint, or skill-adaptive regularization lowers accuracy by 0.7, 1.3, and 1.5 percentage points, respectively. Removing the skill-bonus reward yields 42.8%, while replacing skill-balanced replay with uniform replay yields 42.9%. These results support the contributions of targeted generation and skill-aware learning to the full framework.

Composite Solver reward. Removing trajectory agreement r_{traj} reduces accuracy to 41.9%, while removing the CoT reward r_{cot} reduces it to 42.0%. Combining a frozen Proposer with removal of r_{cot} further lowers accuracy to 40.7%. This combined reduction highlights the value of retaining both adaptive question generation and CoT-quality feedback.

Table 3: Effect of dialogue length and Solver conditioning.

Configuration	**Acc. (%)**
$K_{\text{max}} = 1$ (single turn)	41.9
$K_{\text{max}} = 3$	42.3
$K_{\text{max}} = 5$ (**TerraEvo**)	**43.6**
$K_{\text{max}} = 7$	42.7
Conditioning ablation at $K_{max} = 5$	
w/o Solver-conditioned follow-ups	42.1

Socratic dialogue. Table 3 examines the maximum dialogue length. Accuracy increases from 41.9% with one turn to 42.3% with three turns and peaks at 43.6% with five turns. Increasing the limit to seven turns yields 42.7%, suggesting that additional turns do not necessarily improve learning. With $K_{\text{max}} = 5$, removing Solver-conditioned follow-ups reduces accuracy to 42.1%. This comparison supports the importance of using the Solver's answer disagreement to guide subsequent questions.

CoT reward components. Table 4 progressively adds the components of r_{cot}. Specificity improves accuracy from 42.0% to 42.3%, followed by consistency (42.5%) and length (42.8%). Adding presence completes the reward, reaching 43.6%, a gain of 1.6 percentage points over no CoT reward.

Table 4: Cumulative CoT reward ablations.

Active CoT components	**Acc. (%)**
None	42.0
Specificity	42.3
Specificity, consistency	42.5
Specificity, consistency, length	42.8
All four (**TerraEvo**)	**43.6**

6 Conclusion

We introduced TerraEvo, a self-evolving framework that improves EO vision-language models using unlabeled imagery without additional human annotations or an external judge. TerraEvo combines skill-targeted Socratic dialogue, quality supervision, and adaptive question rewards with composite Solver rewards and knowledge preservation. Across five remote-sensing benchmarks, it improves average accuracy from 51.7% to 54.2%, achieving the highest average among the evaluated models. Ablations support the contributions of Solver-conditioned follow-ups, skill balancing, and complementary reward components. TerraEvo relies on internal verification and a predefined skill taxonomy, which may limit its ability to detect shared prediction errors and to accommodate new task types. Future work will address these limitations through stronger visual verification to catch shared errors and adaptive skill discovery to support new tasks. We also plan to extend TerraEvo to larger models, segmentation tasks, and additional EO modalities such as SAR and multispectral imagery.

Disclosure of AI Use

We used a large language model to improve the manuscript's language and clarity.

References

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025a.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025b.

Xiuwei Chen, Wentao Hu, Hanhui Li, Jun Zhou, Zisheng Chen, Meng Cao, Yihan Zeng, Kui Zhang, Yu-Jie Yuan, Jianhua Han, et al. C2-evo: Co-evolving multimodal data and model for self-improving reasoning. 2025.

Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, et al. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *arXiv preprint arXiv:2412.05271*, 2024a.

Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 24185–24198, 2024b.

Gong Cheng, Junwei Han, and Xiaoqiang Lu. Remote sensing image scene classification: Benchmark and state of the art. *Proceedings of the IEEE*, 105(10):1865–1883, 2017.

Gordon Christie, Neil Fendley, James Wilson, and Ryan Mukherjee. Functional map of the world. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 6172–6180, 2018.

Yunkai Dang, Meiyi Zhu, Donghao Wang, Yizhuo Zhang, Jiacheng Yang, Qi Fan, Yuekun Yang, Wenbin Li, Feng Miao, and Yang Gao. A benchmark for ultra-high-resolution remote sensing mllms. *arXiv preprint arXiv:2512.17319*, 2025.

Muhammad Danish, Muhammad Akhtar Munir, Syed Roshaan Ali Shah, Kartik Kuckreja, Fahad Shahbaz Khan, Paolo Fraccaro, Alexandre Lacoste, and Salman Khan. Geobench-vlm: Benchmarking vision-language models for geospatial tasks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 7132–7142, 2025.

Muhammad Sohail Danish, Muhammad Akhtar Munir, Syed Shah, Muhammad Haris Khan, Rao Anwer, Jorma Laaksonen, Fahad Khan, and Salman Khan. Terrafm: A scalable foundation model for unified multisensor earth observation. In *International Conference on Learning Representations*, volume 2026, pp. 12787–12809, 2026.

Yanghua Di, Zhiguo Jiang, and Haopeng Zhang. A public dataset for fine-grained ship classification in optical remote sensing images. *Remote Sensing*, 13(4):747, 2021.

Jian Ding, Nan Xue, Gui-Song Xia, Xiang Bai, Wen Yang, Michael Ying Yang, Serge Belongie, Jiebo Luo, Mihai Datcu, Marcello Pelillo, et al. Object detection in aerial images: A large-scale benchmark and challenges. *IEEE transactions on pattern analysis and machine intelligence*, 44 (11):7778–7796, 2021.

Mustansar Fiaz, Hiyam Debary, Paolo Fraccaro, Danda Paudel, Luc Van Gool, Fahad Khan, and Salman Khan. Geovlm-r1: Reinforcement fine-tuning for improved remote sensing reasoning. *arXiv preprint arXiv:2509.25026*, 2025a.

Mustansar Fiaz, Mubashir Noman, Hiyam Debary, Kamran Ali, and Hisham Cholakkal. Hyretchange: A hybrid retentive network for remote sensing change detection. In *IGARSS 2025-2025 IEEE International Geoscience and Remote Sensing Symposium*, pp. 7209–7213. IEEE, 2025b.

Ritwik Gupta, Bryce Goodman, Nirav Patel, Ricky Hosfelt, Sandra Sajeev, Eric Heim, Jigar Doshi, Keane Lucas, Howie Choset, and Matthew Gaston. Creating xbd: A dataset for assessing building damage from satellite imagery. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pp. 10–17, 2019a.

Ritwik Gupta, Richard Hosfelt, Sandra Sajeev, Nirav Patel, Bryce Goodman, Jigar Doshi, Eric Heim, Howie Choset, and Matthew Gaston. xbd: A dataset for assessing building damage from satellite imagery. *arXiv preprint arXiv:1911.09296*, 2019b.

Yicheng He, Chengsong Huang, Zongxia Li, Jiaxin Huang, and Yonghui Yang. Visplay: Self-evolving vision-language models from images. *arXiv preprint arXiv:2511.15661*, 2025.

Yicheng He, Chengsong Huang, Zongxia Li, Jiaxin Huang, and Yonghui Yang. Visplay: Self-evolving vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 26274–26284, 2026.

Chengsong Huang, Wenhao Yu, Xiaoyang Wang, Hongming Zhang, Zongxia Li, Ruosen Li, Jiaxin Huang, Haitao Mi, and Dong Yu. R-zero: Self-evolving reasoning llm from zero data. In *International Conference on Learning Representations*, volume 2026, pp. 130770–130790, 2026.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024.

Jeremy Irvin, Emily Liu, Joyce Chen, Ines Dormoy, Jinyoung Kim, Samar Khanna, Zhuo Zheng, and Stefano Ermon. Teochat: A large vision-language assistant for temporal earth observation data. In *International Conference on Learning Representations*, volume 2025, pp. 68883–68911, 2025.

Kartik Kuckreja, Muhammad Sohail Danish, Muzammal Naseer, Abhijit Das, Salman Khan, and Fahad Shahbaz Khan. Geochat: Grounded large vision-language model for remote sensing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 27831–27840, 2024.

Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, et al. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*, 2024.

Ke Li, Gang Wan, Gong Cheng, Liqiu Meng, and Junwei Han. Object detection in optical remote sensing images: A survey and a new benchmark. *ISPRS journal of photogrammetry and remote sensing*, 159:296–307, 2020.

Tingyu Li, Zheng Sun, Jingxuan Wei, Conghui He, Lijun Wu, and Cheng Tan. Decouple to generalize: Context-first self-evolving learning for data-scarce vision-language reasoning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 29357–29366, 2026a.

Zongxia Li, Hongyang Du, Chengsong Huang, Xiyang Wu, Lantao Yu, Yicheng He, Jing Xie, Xiaomin Wu, Zhichao Liu, Jiarui Zhang, et al. Mm-zero: Self-evolving multi-model vision language models from zero data. *arXiv preprint arXiv:2603.09206*, 2026b.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 26286–26296. IEEE, 2024.

Jiaqi Liu, Lang Sun, Ronghao Fu, and Bo Yang. Towards faithful reasoning in remote sensing: A perceptually-grounded geospatial chain-of-thought for vision-language models. In *International Conference on Learning Representations*, volume 2026, pp. 6535–6563, 2026a.

Ruixun Liu, Bowen Fu, Jiayi Song, Kaiyu Li, Wanchen Li, Lanxuan Xue, Hui Qiao, Weizhan Zhang, Deyu Meng, and Xiangyong Cao. Zoomearth: Active perception for ultra-high-resolution geospatial vision-language tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 34877–34888, 2026b.

Yexiang Liu, Jie Cao, Zekun Li, Ran He, and Tieniu Tan. Breaking mental set to improve reasoning through diverse multi-agent debate. In *International Conference on Learning Representations*, volume 2025, pp. 24666–24725, 2025.

Sylvain Lobry, Diego Marcos, Jesse Murray, and Devis Tuia. Rsvqa: Visual question answering for remote sensing data. *IEEE Transactions on Geoscience and Remote Sensing*, 58(12):8555–8566, 2020.

Junwei Luo, Zhen Pang, Yongjun Zhang, Tingzhu Wang, Linlin Wang, Bo Dang, Jiangwei Lao, Jian Wang, Jingdong Chen, Yihua Tan, et al. Skysensegpt: A fine-grained instruction tuning dataset and model for remote sensing vision-language understanding. *arXiv preprint arXiv:2406.10100*, 2024.

Dilxat Muhtar, Zhenshi Li, Feng Gu, Xueliang Zhang, and Pengfeng Xiao. Lhrs-bot: Empowering remote sensing with vgi-enhanced large multimodal language model. In *European Conference on Computer Vision*, pp. 440–457. Springer, 2024.

Chao Pang, Xingxing Weng, Jiang Wu, Jiayu Li, Yi Liu, Jiaxing Sun, Weijia Li, Shuai Wang, Litong Feng, Gui-Song Xia, et al. Vhm: Versatile and honest vision language model for remote sensing image analysis. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 6381–6388, 2025.

Lin Qiu, Hanqing Zeng, Yao Liu, Bingjun Sun, Guangdeng Liao, and Ji Liu. Duel: Adversarial self-play for multimodal reasoning. *arXiv preprint arXiv:2605.24794*, 2026.

Aaditya Singh, Adam Fry, Adam Perelman, Adam Tart, Adi Ganesh, Ahmed El-Kishky, Aidan McLaughlin, Aiden Low, AJ Ostrow, Akhila Ananthram, et al. Openai gpt-5 system card. *arXiv preprint arXiv:2601.03267*, 2025.

Sagar Soni, Akshay Dudhane, Hiyam Debary, Mustansar Fiaz, Muhammad Akhtar Munir, Muhammad Sohail Danish, Paolo Fraccaro, Campbell D Watson, Levente J Klein, Fahad Shahbaz Khan, 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.

Xian Sun, Peijin Wang, Zhiyuan Yan, Feng Xu, Ruiping Wang, Wenhui Diao, Jin Chen, Jihao Li, Yingchao Feng, Tao Xu, et al. Fair1m: A benchmark dataset for fine-grained object recognition in high-resolution remote sensing imagery. *ISPRS Journal of Photogrammetry and Remote Sensing*, 184:116–130, 2022.

Meghana Sunil, Manikandarajan Venmathimaran, and Muthu Subash Kavitha. ireasoner: Trajectory-aware intrinsic reasoning supervision for self-evolving large multimodal models. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 29361–29372, 2026.

Omkar Thawakar, Shravan Venkatraman, Ritesh Thawkar, Abdelrahman Shaker, Hisham Cholakkal, Rao Muhammad Anwer, Salman Khan, and Fahad Khan. Evolmm: Self-evolving large multimodal models with continuous rewards. *arXiv preprint arXiv:2511.16672*, 2025.

Fengxiang Wang, Hongzhen Wang, Zonghao Guo, Di Wang, Yulin Wang, Mingshuo Chen, Qiang Ma, Long Lan, Wenjing Yang, Jing Zhang, et al. Xlrs-bench: Could your multimodal llms understand extremely large ultra-high-resolution remote sensing imagery? In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14325–14336. IEEE, 2025a.

Fengxiang Wang, Mingshuo Chen, Yueying Li, Di Wang, Haotian Wang, Zonghao Guo, Zefan Wang, Shan Boqi, Long Lan, Yulin Wang, et al. Geollava-8k: Scaling remote-sensing multimodal large language models to 8k resolution. *Advances in Neural Information Processing Systems*, 38: 159185–159218, 2026a.

Fengxiang Wang, Mingshuo Chen, Yueying Li, Yajie Yang, Yifan Zhang, Long Lan, Xue Yang, Hongda Sun, Yulin Wang, Di Wang, et al. Geoeyes: On-demand visual focusing for evidence-grounded understanding of ultra-high-resolution remote sensing imagery. *arXiv preprint arXiv:2602.14201*, 2026b.

Qinsi Wang, Bo Liu, Tianyi Zhou, Jing Shi, Yueqian Lin, Yiran Chen, Hai Li, Kun Wan, and Wentian Zhao. Vision-zero: Scalable vlm self-evolution via multi-agent self-play. In *International Conference on Learning Representations*, volume 2026, pp. 106716–106744, 2026c.

Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, et al. Internvl3. 5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *arXiv preprint arXiv:2508.18265*, 2025b.

Tianhao Wu, Weizhe Yuan, Olga Golovneva, Jing Xu, Yuandong Tian, Jiantao Jiao, Jason E Weston, and Sainbayar Sukhbaatar. Meta-rewarding language models: Self-improving alignment with llm-as-a-meta-judge. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 11548–11565, 2025.

Ziyi Yang, Weizhou Shen, Chenliang Li, Ruijun Chen, Fanqi Wan, Ming Yan, Xiaojun Quan, and Fei Huang. Spell: Self-play reinforcement learning for evolving long-context language models. In *International Conference on Learning Representations*, volume 2026, pp. 59910–59942, 2026.

Liang Yao, Fan Liu, Hongbo Lu, Chuanyi Zhang, Rui Min, Shengxiang Xu, Shimin Di, and Pai Peng. Remotereasoner: Towards unifying geospatial reasoning workflow. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 11883–11891, 2026.

Yang Zhan, Zhitong Xiong, and Yuan Yuan. Rsvg: Exploring data and models for visual grounding on remote sensing data. *IEEE Transactions on Geoscience and Remote Sensing*, 61:1–13, 2023.

Wei Zhang, Miaoxin Cai, Tong Zhang, Yin Zhuang, and Xuerui Mao. Earthgpt: A universal multimodal large language model for multi-sensor image comprehension in remote sensing domain. *IEEE Transactions on Geoscience and Remote Sensing*, 2024.

YiFan Zhang, Huanyu Zhang, Haochen Tian, Chaoyou Fu, Shuangqing Zhang, Junfei Wu, Feng Li, Kun Wang, Qingsong Wen, Zhang Zhang, et al. Mme-realworld: Could your multimodal llm challenge high-resolution real-world scenarios that are difficult for humans? In *International Conference on Learning Representations*, volume 2025, pp. 89655–89701, 2025.

Zilun Zhang, Zian Guan, Tiancheng Zhao, Haozhan Shen, Yuxiang Cai, Zhonggen Su, Yongheng Shang, Zhaojun Liu, Jianwei Yin, and Xiang Li. Geo-r1: Improving few-shot geospatial referring expression understanding with reinforcement fine-tuning. *ISPRS Journal of Photogrammetry and Remote Sensing*, 237:113–129, 2026.

Andrew Zhao, Yiran Wu, Tong Wu, Quentin Xu, Yang Yue, Matthieu Lin, Shenzhi Wang, Qingyun Wu, Zilong Zheng, and Gao Huang. Absolute zero: Reinforced self-play reasoning with zero data. *Advances in Neural Information Processing Systems*, 38:105816–105879, 2026.

Weixun Zhou, Shawn Newsam, Congmin Li, and Zhenfeng Shao. Patternnet: A benchmark dataset for performance evaluation of remote sensing image retrieval. *ISPRS journal of photogrammetry and remote sensing*, 145:197–209, 2018.

Jiashun Zhu, Ronghao Fu, Jiasen Hu, Jing Huang, Nachuan Xing, and Bo Yang. Geovista: Visually grounded active perception for vision-language understanding of ultra-high-resolution remote sensing images. *arXiv preprint arXiv:2605.14475*, 2026.

Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, et al. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*, 2025.

A APPENDIX

A.1 IMPLEMENTATION DETAILS

Both the Proposer and Solver are initialized from Qwen3-VL-4B-Instruct (Bai et al., 2025a), using separate LoRA adapters on a shared backbone. The adapters use rank 16, $\alpha_{\text{LoRA}} = 32$, and dropout 0.05 across all attention and MLP projections, while the vision encoder remains frozen. We train for 3K steps on 4 NVIDIA A100 80GB GPUs in bfloat16 precision using AdamW with a learning rate of 10^{-6}, weight decay of 0.01, and gradient-norm clipping at 1.0. The Solver generates $M = 14$ responses per question. The Socratic chain runs for up to $K_{\max} = 5$ turns and stops when answer agreement reaches $\tau_{\text{stop}} = 0.70$ (Section 4.2). The Proposer is updated every $F = 2$ training steps.

The Solver's base reference-regularization coefficient is initialized to $\beta_c = 0.02$ and adjusted at rate $\eta = 0.08$ toward a target of 0.02. The skill tracker maintains $W = 500$ resolved labels, with hard-forcing activated at $\tau_{\text{force}} = 0.40$ and up to $R_{\max} = 3$ total generation attempts per turn. The confidence gate uses a base lower threshold of $\tau_{\min}^{\text{base}} = 0.45$ and an upper threshold of $\tau_{\max} = 0.75$. Quality supervision activates after $S_{\text{warm}} = 200$ steps, and the skill and validity bonuses share weight $\lambda = 0.40$.

The Proposer's difficulty reward uses $\alpha_d = 0.60$, $\alpha_e = 0.40$, and $\alpha_b = 0.10$, with target entropy $\mu_H = 1.60$ and bandwidth $\sigma_H = 0.45$. The agreement moving average and the Proposer baseline use momentum $m_e = m_b = 0.90$, with reference offset $\delta = 0.12$, while a BLEU overlap threshold of 0.50 controls the repetition penalty. The Solver's soft answer-agreement reward uses exponent $\gamma = 0.60$ and a reasoning-length penalty with weight $w_\ell = 0.05$ and target $L_{\text{tgt}} = 150$ words. The trajectory and CoT reward weights use warmup fractions of 10% and 5%, respectively, followed by linear ramps toward $w_{\text{step}} = 0.70$ and $w_{\text{cot}} = 0.30$. Solver and Proposer generations are capped at 512 and 256 tokens, respectively.

A.2 REWARD COMPUTATION AND QUALITY CONTROL

We detail the trajectory reward, verification prompts, and adaptive replay and regularization mechanisms used in TerraEvo.

A.2.1 TRAJECTORY AGREEMENT

Trajectory agreement measures reasoning similarity across responses sharing the majority answer. It complements soft answer agreement by distinguishing responses with the same answer according to their reasoning passages.

Reasoning representation. Each response's reasoning passage is represented by an embedding $\mathbf{h}_j$. Let $\mathcal{D}$ denote the majority-answer group among the M sampled responses, and let $\mathcal{D}_v \subseteq \mathcal{D}$ contain its responses with nonempty reasoning.

Group center and similarity. When $|\mathcal{D}_v| \geq N_{\min}$, we average the reasoning embeddings and normalize the result to obtain the group center

$$\bar{\mathbf{h}} = \frac{1}{|\mathcal{D}_v|} \sum_{i \in \mathcal{D}_v} \mathbf{h}_i, \qquad \mathbf{g} = \frac{\bar{\mathbf{h}}}{\|\bar{\mathbf{h}}\|_2 + \varepsilon}, \tag{12}$$

where $\varepsilon = 10^{-8}$ ensures numerical stability. Each response contributes to the center against which it is compared. Its similarity and the corresponding score in $[0, 1]$ are

$$\kappa_j = \text{clip}\left(\frac{\mathbf{h}_j^\top \mathbf{g}}{\|\mathbf{h}_j\|_2 + \varepsilon}, -1, 1 \right), \qquad s_j = \frac{\kappa_j + 1}{2}. \tag{13}$$

Higher scores indicate closer agreement with the group's mean reasoning representation.

Reward computation. The trajectory reward combines reasoning similarity with the majority group's proportion $d = |\mathcal{D}|/M$,

$$r_{\text{traj}}^{(j)} = \begin{cases} d^{\gamma_d} s_j, & j \in \mathcal{D}_v,\ |\mathcal{D}_v| \geq N_{\min}, \\ 0, & \text{otherwise}, \end{cases} \tag{14}$$

where $N_{\text{min}} = 2$ is the minimum comparison-group size and $\gamma_d = 0.7$ controls density weighting. Responses outside the majority group or without reasoning receive zero trajectory reward. All trajectory rewards are zero when fewer than two eligible passages are available.

Design rationale. Reasoning similarity provides a consistency signal beyond final-answer agreement. The minimum group size prevents comparison against a single response, while density weighting accounts for the proportion supporting the majority answer. These signals capture shared reasoning patterns without independently establishing visual correctness.

Trajectory rewards are computed over the original M responses before filtering. The composite reward then determines the highest-reward $G = M - \lfloor M/2 \rfloor$ responses retained for the Solver update.

A.2.2 Verification Prompts

The Solver uses the input images to assess question validity and majority-answer support separately. Question validation produces the signal ρ used in the Proposer reward. Answer verification provides an additional condition for replay admission but does not contribute to the Proposer reward.

Question validation. The following prompt assesses visual grounding, semantic validity, answerability, and consistency with the resolved skill.

```
You are a strict visual QA quality judge for remote sensing
images.
Answer YES or NO only.
Answer YES only if ALL hold:
  1. The question is grounded in the image(s) content.
  2. The question is semantically well-formed.
  3. The question can be answered from the image(s) alone.
  4. The question matches the declared skill category.
Valid skills: {skills}
Skill: {skill}
Question: {question}
YES or NO:
```

The placeholders {skill} and {skills} denote the resolved skill and supported taxonomy, respectively.

Majority-answer verification. For questions passing validation and the confidence gate, the following prompt checks whether the majority answer is supported by the image.

```
You are a visual QA answer verifier for remote sensing
images.
Answer YES or NO only.
Answer YES if the candidate answer is factually correct and
supported by the image(s).
Question: {question}
Candidate answer: {answer}
YES or NO:
```

The candidate answer is the majority answer among the sampled responses, and an affirmative verification output indicates a passed check. Both checks are bypassed during the first $S_{\text{warm}} = 200$ steps. These internal assessments support filtering without independently guaranteeing correctness.

A.2.3 Adaptive Replay and Regularization

Skill-specific confidence gate. Replay admission accounts for the Solver's recent answer agreement within each skill. For skill s, $\mathcal{H}_s$ stores agreement values from its latest 50 dialogue turns, including the current turn. We estimate recent agreement as

$$\bar{u}_s = \begin{cases} 0.5, & |\mathcal{H}_s| < 5, \\ \dfrac{1}{|\mathcal{H}_s|} \displaystyle\sum_{u \in \mathcal{H}_s} u, & \text{otherwise.} \end{cases} \tag{15}$$

A default value of 0.5 is used until five observations are available. The lower admission threshold is

$$\tau_{\text{min}}(s) = \begin{cases} \min\big(0.70, \tau_{\text{min}}^{\text{base}} + 0.40\big), & \bar{u}_s < 0.45, \\ \tau_{\text{min}}^{\text{base}}, & \text{otherwise}, \end{cases} \tag{16}$$

where $\tau_{\text{min}}^{\text{base}} = 0.45$. Thus, skills with low recent agreement receive a stricter lower threshold of 0.70. The upper threshold remains fixed at $\tau_{\text{max}} = 0.75$.

After warmup, replay admission requires both verification checks to pass and answer agreement to fall within the accepted range,

$$\rho(x, q, \hat{s}) = 1, \qquad \nu(x, q, \hat{a}) = 1, \qquad \tau_{\text{min}}(\hat{s}) \le u(x, q) \le \tau_{\text{max}}, \tag{17}$$

where $\hat{s}$ is the resolved skill and ν denotes majority-answer verification. These conditions are bypassed during warmup. The adaptive lower bound requires stronger agreement for skills with low recent consistency, while the upper bound excludes highly consensual questions.

Skill-balanced replay. Each skill buffer stores up to 50 retained responses and their composite rewards. Every F training steps, we sample one response from each nonempty buffer and skip entries whose Solver prompt matches the current prompt. Each remaining response receives a separate Solver update, with its stored reward used directly as the advantage. This maintains exposure across skills without generating new responses or recomputing rewards.

Adaptive reference regularization. For a response assigned to skill s, the Solver's regularization coefficient is

$$\beta(s) = \beta_c \left(1 + 2b_{\text{skill}}(s)\right). \tag{18}$$

Both regular and replay updates use the current base coefficient and skill bonus. Under-represented skills therefore receive stronger regularization toward the reference policy.

The base coefficient is adjusted according to the difference between the measured mean KL divergence $\widehat{D}$ and its target D_{tgt},

$$z = \text{clip}\left(\eta \frac{\widehat{D} - D_{\text{tgt}}}{\max(D_{\text{tgt}}, 10^{-8})}, -20, 20\right), \tag{19}$$

$$\beta_c \leftarrow \text{clip}\big(\beta_c e^z, 10^{-8}, 10^2\big), \tag{20}$$

where $\eta = 0.08$ and $D_{\text{tgt}} = 0.02$. The coefficient increases when the measured statistic exceeds the target and decreases when it falls below it. Clipping bounds the adjustment and the resulting coefficient. The Proposer adapts its own reference coefficient β_p independently using the same rule, without skill scaling.

A.3 Training Algorithm

Algorithm 1 summarizes the interaction and learning process within one training step of TerraEvo. An under-represented skill guides question generation, while feedback from the Solver shapes subsequent questions through multi-turn Socratic dialogue. Quality supervision and the Proposer reward support selection of the highest-reward question, whose sampled responses are scored to train the Solver. Eligible responses are stored for skill-balanced replay. Every F steps, replay updates are followed by a Proposer update using the selected question's reward, allowing both roles to improve through self-generated feedback.

A.4 Analysis

A.4.1 Effect of Proposer-Side Skill Guidance

Figure 3 compares TerraEvo with a generic Proposer without skill guidance. Removing skill guidance reduces GeoBench-VLM accuracy from 43.6% to 41.8% and leads to a more uneven question distribution. Across steps 400–3,000, the generic Proposer achieves a mean normalized skill entropy of 0.727, compared with 0.937 for TerraEvo. Its questions are dominated by counting, with limited coverage of assessment, temporal, and reasoning skills. In contrast, TerraEvo sustains broader, more balanced coverage across all eight skills.

Algorithm 1 One Training Step of TerraEvo

Require: Step t, Solver π_s, Proposer π_q, skill tracker, replay buffers, update interval F

Sample image record x; initialize feedback $f_0 = \varnothing$
Select $s^{\text{tgt}} \leftarrow \arg\max_{s \in \mathcal{S}} b_{\text{skill}}(s)$ {Eq. 4}
Fix force $\leftarrow [b_{\text{skill}}(s^{\text{tgt}}) > \tau_{\text{force}}]$
for $k = 1$ to $K_{\max}$ **do**
  Generate q_k using x, the target skill, and feedback f_{k-1}
  If forced, allow up to $R_{\max}$ total generation attempts
  Resolve skill $\hat{s}_k$ and update the skill window
  Sample M Solver responses; compute agreement u_k and entropy
  Update skill agreement history; assess validity and replay eligibility {Section 4.3.1}
  Compute question reward $r_q^{(k)}$ {Eq. 6}
  if $u_k \geq \tau_{\text{stop}}$ **then**
    break
  end if
  Build feedback f_k from majority and minority answers
end for
Select $k^\star = \arg\max_{k \in \{1,\ldots,K\}} r_q^{(k)}$ over completed turns; set $q^\star = q_{k^\star}$
Score $q^\star$'s existing responses; retain the highest-reward half {Eq. 9}
Store eligible responses and rewards in the corresponding skill buffer
Update π_s with skill-adaptive regularization {Eqs. 2, 11}
if $t \bmod F = 0$ **then**
  for each skill with a non-empty replay buffer **do**
    Sample one item; skip if its Solver prompt matches the current prompt
    Update π_s using the stored reward and the item's skill-adaptive coefficient {Eq. 11}
  end for
  Update π_q using the selected Proposer output and reward $r_q^{(k^\star)}$; update baseline b {Eq. 3}
end if

A.4.2 Skill-Targeted Socratic Follow-ups

TerraEvo selects the most under-represented skill (§4.1) and directs the Proposer to generate questions within that category. Figure 4 illustrates this interaction for spatial relations. With Solver-conditioned feedback (left), each follow-up drills deeper into the object introduced by the previous turn rather than moving on, progressively narrowing a broad initial reference into one the Solver can answer precisely. Without this feedback (right), each question targets a different object in the scene, so nothing raised in an earlier turn is ever revisited or further localized. Similarly, Figure 5 shows an analogous example for counting: Solver-conditioned feedback (left) isolates the source of disagreement and reconciles it into a corrected count, while questioning without this feedback (right) drifts to unrelated scene attributes and leaves the miscount uncorrected.

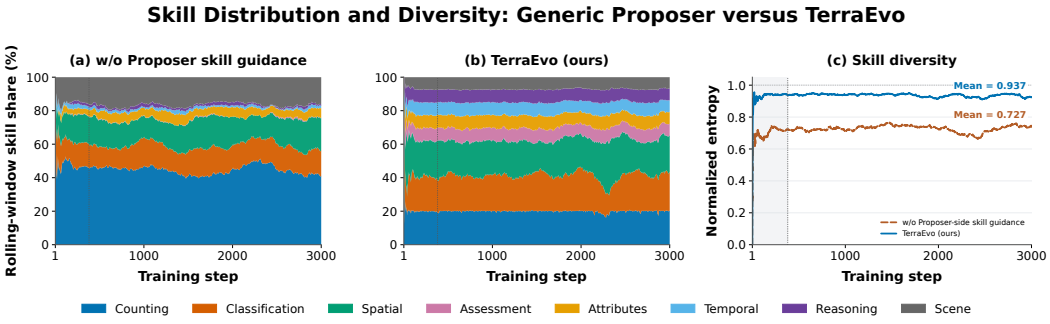


Figure 3: Effect of generic Proposer without skill guidance on question diversity. (a) and (b) show skill proportions in a rolling window. (c) compares normalized skill entropy, where higher values indicate more even coverage.

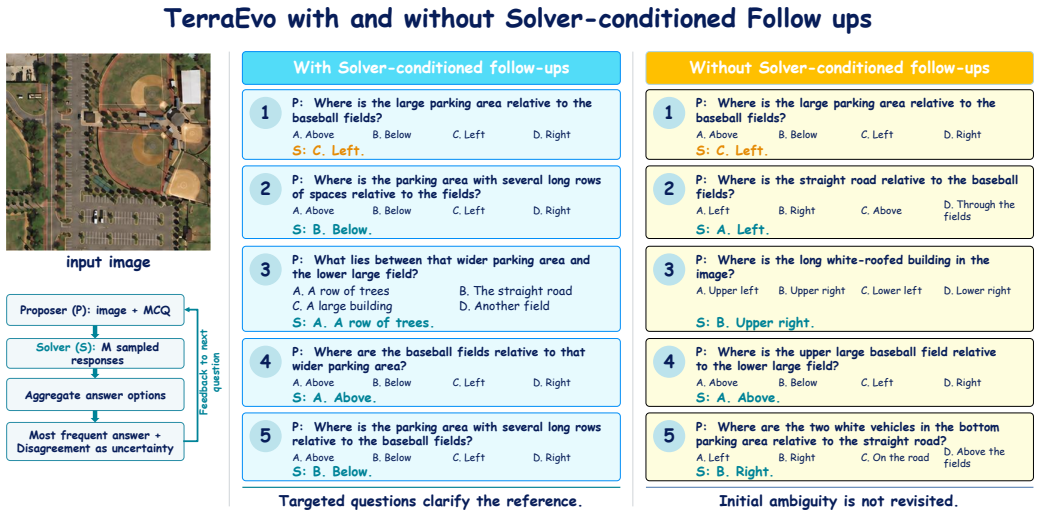


Figure 4: **Illustrative comparison of spatial skill question generation with and without Solver-conditioned follow-ups.** With conditioning (left), the Proposer uses Solver feedback to clarify which parking area the question refers to and check its spatial relation to the fields. Without conditioning (right), subsequent questions address other spatial relations, leaving the initial ambiguity unresolved.

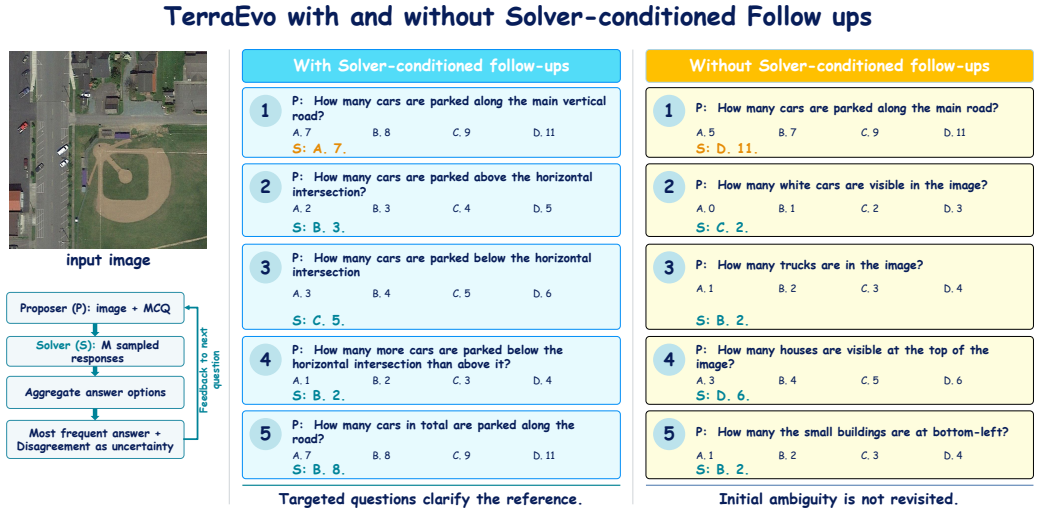


Figure 5: **Illustrative comparison of counting-skill question generation with and without Solver-conditioned follow-ups.** With conditioning (left), the Proposer uses Solver feedback to isolate where the initial disagreement comes from and reconcile the sub-counts into a corrected total. Without conditioning (right), subsequent questions address other scene attributes, leaving the initial miscount uncorrected.

A.5 Quality Audit of Self-Generated Training Data

Metrics. *Answer correct*: the Solver's final answer matches the answer verified by a human judge. *Valid question*: the generated question is well-posed, unambiguous, and has a single defensible answer. *Responses agree*: independent Solver responses to the same question describe the same visible content. *No unsupported reasoning detected*: no response describes content that contradicts the other responses or refers to an image that was not provided. *Reasoning consistent with answer*: the reasoning leads to the final answer given in the response.

Table 5: Quality audit of randomly 100 generated questions and their Solver responses, sampled from the 3K training steps. †Verified by a human judge; all other metrics are assessed by an external model. *Computed at the response level.

Metric	Rate (%)
Answer correct†	76
Valid question	77
Responses to the same question agree	93
No unsupported reasoning detected	82
Reasoning consistent with final answer*	88

Results. As shown in Table 5, 77% of generated questions are valid, and a human judge verifies that the Solver answers 76% of the questions correctly. The responses also exhibit substantial internal consistency such that responses to the same question agree for 93% of questions, no unsupported reasoning is detected for 82% of questions, and 88% of individual responses give reasoning consistent with their own final answer.

A.6 Detailed Comparison on Five Challenging RS Benchmarks

Detailed per-task results for each benchmark are shown in Tables 6, 7, 8, 9, and 10 for GeoBench-VLM (Danish et al., 2025), LHRS-Bench (Muhtar et al., 2024), MME-RealWorld-RS (Zhang et al., 2025), RSHR-Bench (Dang et al., 2025), and XLRS-Bench (Wang et al., 2025a), respectively. Overall, **TerraEvo** (Ours) method achieves the best average score across all five benchmarks (54.2%) compared to general-purpose and specialized remote-sensing MLLMs.

Table 6: Per-task zero-shot accuracy (%) on GeoBench-VLM for closed-source, open-source, and remote-sensing MLLMs, compared against our self-evolving framework. Here AC = Aircraft Classification, BC = Building Counting, CC = Crop Classification, DC = Disaster Classification, FRA = Fire Risk Assessment, GAC = General Aircraft Counting, GVC = General Vehicle Counting, LUC = Land Use Classification, MDC = Marine Debris Counting, ScC = Scene Classification, ShC = Ship Classification, SAC = Specific Aircraft Counting, SVC = Specific Vehicle Counting, SR = Spatial Relationship, THA = Tree Health Assessment, TC = Trees Counting, WBC = Water Bodies Counting, Avg = Average (unweighted mean over all 17 tasks). Bold indicates the best result in each column, including ties.

Model	AC	BC	CC	DC	FRA	GAC	GVC	LUC	MDC	ScC	ShC	SAC	SVC	SR	THA	TC	WBC	Avg
Closed-Source MLLMs																		
GPT-4o (Hurst et al., 2024)	**64.0**	28.2	23.6	**61.7**	20.0	36.0	13.3	62.7	16.0	85.8	56.0	32.1	25.9	**74.7**	13.4	29.4	56.5	41.1
GPT-5 (Singh et al., 2025)	53.0	35.3	23.6	61.2	16.7	37.0	16.0	**66.6**	17.0	87.2	**69.6**	32.1	24.6	70.2	28.7	27.1	60.0	42.7
Open-Source MLLMs																		
InternVL2.5-8B (Chen et al., 2024a)	47.0	24.7	25.4	50.7	20.0	41.0	23.3	53.7	27.0	80.0	29.9	34.3	34.8	67.2	7.0	22.4	52.9	37.7
InternVL3.5-8B (Wang et al., 2025b)	47.0	24.1	18.2	38.3	18.3	**53.0**	38.7	50.0	37.0	**88.4**	31.9	30.7	33.0	33.7	17.8	29.4	55.3	37.9
InternVL3-8B (Zhu et al., 2025)	45.0	24.7	27.3	46.7	19.2	39.0	38.7	55.1	47.0	79.5	36.7	36.4	35.7	64.4	14.6	29.4	47.1	40.4
LLaVA1.5-7B (Liu et al., 2024)	44.0	24.1	20.0	38.3	21.7	23.0	12.0	49.3	16.0	70.4	27.5	17.1	16.5	47.9	16.6	23.5	32.9	29.5
LLaVA-OneVision-7B (Li et al., 2024)	53.0	21.2	23.6	44.9	18.3	27.0	30.0	55.9	30.0	82.7	42.0	23.6	26.3	66.7	8.9	27.1	38.8	36.5
Qwen2.5-VL-3B (Bai et al., 2025b)	45.0	24.7	16.4	47.1	20.8	24.0	31.3	44.1	40.0	70.6	31.9	25.0	31.2	63.1	15.3	28.2	60.0	36.4
Qwen2.5-VL-7B (Bai et al., 2025b)	56.0	**37.6**	20.0	40.1	16.7	40.0	**44.0**	54.1	41.0	80.2	39.6	26.4	33.9	68.5	24.2	24.7	62.4	41.7
Qwen3-VL-4B (Bai et al., 2025a)	48.0	25.9	21.8	56.0	24.2	41.0	20.7	57.3	24.0	79.3	45.9	28.6	26.3	73.4	12.1	21.2	62.4	39.3
Qwen3-VL-8B (Bai et al., 2025a)	50.0	27.7	20.0	52.4	25.8	49.0	37.3	58.3	35.0	81.2	43.5	40.7	28.1	71.7	14.0	20.0	**63.5**	42.2
Domain-Specific Remote Sensing MLLMs																		
GeoChat (Kuckreja et al., 2024)	20.0	17.1	14.5	26.0	17.5	24.0	13.3	19.3	20.0	22.4	17.4	18.6	20.5	25.5	17.2	20.0	17.6	19.5
EarthDial (Soni et al., 2025)	36.0	20.6	**29.1**	18.1	17.5	29.0	18.0	56.3	14.0	77.3	22.7	27.9	28.1	70.6	14.0	24.7	45.9	32.3
VHM (Pang et al., 2025)	46.0	20.0	20.0	38.3	**28.3**	18.0	13.3	60.5	27.0	83.4	29.5	23.6	19.2	63.3	11.5	22.4	48.2	33.7
GeoLLaVA-8K (Wang et al., 2026a)	28.0	21.2	25.4	21.6	24.2	33.0	36.7	23.2	**51.0**	30.4	21.3	40.7	37.5	26.2	**32.5**	**30.6**	22.4	29.7
ZoomEarth (Liu et al., 2026b)	42.0	17.1	23.6	40.5	15.8	29.0	29.3	39.3	24.0	68.2	31.4	28.6	27.7	52.4	15.3	25.9	23.5	31.4
Geoeyes (Wang et al., 2026b)	38.0	25.3	**29.1**	35.2	14.2	33.0	22.7	51.5	35.0	80.5	30.9	30.0	28.1	55.1	19.1	21.2	56.5	35.6
GeoVista (Zhu et al., 2026)	51.0	22.3	23.6	34.8	15.8	46.5	19.1	53.7	47.8	85.3	35.3	27.1	26.1	65.7	26.1	21.2	51.8	38.4
Self-Evolving																		
TerraEvo (Ours)	49.0	26.5	23.6	58.6	20.0	**53.0**	34.7	58.5	31.0	80.0	46.4	**42.9**	**38.8**	71.5	18.5	**30.6**	57.6	**43.6**

Table 7: LHRS-Bench per-class accuracy (%) for closed-source, open-source, and remote-sensing MLLMs. Bold indicates the best result in each column, including ties.

Model	Identity	Color	Orientation	Shape	Quantity	Area	Distance	Resolution	Modality	Location	Reasoning	Average
Closed-Source MLLMs												
GPT-4o (Hurst et al., 2024)	67.8	61.1	41.0	81.1	41.6	69.3	36.4	**90.5**	95.7	66.7	**84.8**	66.9
GPT-5 (Singh et al., 2025)	**77.4**	77.0	**64.1**	86.5	**60.6**	73.3	40.9	76.2	**100.0**	77.5	82.6	**74.2**
Open-Source MLLMs												
InternVL2.5-8B (Chen et al., 2024a)	70.3	77.0	28.2	75.7	51.1	68.0	59.1	57.1	34.8	71.6	82.6	61.4
InternVL3.5-8B (Wang et al., 2025b)	71.8	71.7	33.3	73.0	53.3	69.3	50.0	42.9	26.1	79.4	80.4	59.2
InternVL3-8B (Zhu et al., 2025)	72.4	73.5	38.5	73.0	56.9	72.0	59.1	52.4	26.1	76.5	80.4	61.9
LLaVA1.5-7B (Liu et al., 2024)	52.8	48.7	20.5	59.5	35.0	44.0	40.9	14.3	13.0	58.3	76.1	42.1
LLaVA-OneVision-7B (Li et al., 2024)	72.1	69.9	43.6	75.7	58.4	66.7	45.5	9.5	95.7	75.5	80.4	63.0
Qwen2.5-VL-3B (Bai et al., 2025b)	68.0	74.3	25.6	81.1	41.6	62.7	**68.2**	52.4	47.8	74.0	78.3	61.3
Qwen2.5-VL-7B (Bai et al., 2025b)	71.0	70.8	38.5	83.8	47.5	73.3	54.5	4.8	43.5	76.0	**84.8**	58.9
Qwen3-VL-4B (Bai et al., 2025a)	74.6	**79.7**	28.2	86.5	49.6	**78.7**	50.0	42.9	47.8	82.8	**84.8**	64.1
Qwen3-VL-8B (Bai et al., 2025a)	74.8	**79.7**	30.8	81.1	55.5	69.3	50.0	38.1	43.5	80.9	82.6	62.4
Domain-Specific Remote Sensing MLLMs												
GeoChat (Kuckreja et al., 2024)	39.3	35.4	25.6	56.8	27.0	42.7	27.3	4.8	0.0	41.7	41.3	31.1
EarthDial (Soni et al., 2025)	61.2	60.2	25.6	64.9	46.7	61.3	40.9	38.1	26.1	69.6	69.6	51.3
VHM (Pang et al., 2025)	54.4	53.1	28.2	73.0	40.1	53.3	22.7	9.5	26.1	57.8	78.3	45.1
GeoLLaVA-8K (Wang et al., 2026a)	37.1	31.9	25.6	62.2	34.3	28.0	22.7	9.5	73.9	38.2	41.3	36.8
ZoomEarth (Liu et al., 2026b)	51.9	60.2	15.4	54.0	44.5	60.0	45.5	52.4	30.4	48.5	58.7	47.4
Geoeyes (Wang et al., 2026b)	71.9	72.6	46.1	70.3	53.3	70.7	59.1	38.1	39.1	76.5	82.6	61.8
GeoVista (Zhu et al., 2026)	66.4	65.5	38.5	70.3	46.0	69.3	36.4	38.1	26.1	72.5	80.4	55.4
Self-Evolving												
TerraEvo (Ours)	76.8	78.8	33.3	**89.2**	54.0	76.0	54.5	38.1	56.5	**84.8**	**84.8**	66.1

Table 8: MME-RealWorld-RS per-category accuracy (%). Mean is the unweighted average over the three categories. Bold indicates the best result in each column, including ties.

Model	Position	Color	Count	Mean
Closed-Source MLLMs				
GPT-4o (Hurst et al., 2024)	64.8	53.5	30.9	49.7
GPT-5 (Singh et al., 2025)	70.2	61.5	30.9	54.2
Open-Source MLLMs				
InternVL2.5-8B	57.8	56.7	27.1	47.2
InternVL3.5-8B	64.3	53.7	24.6	47.5
InternVL3-8B	66.7	62.8	23.2	50.9
LLaVA1.5-7B	26.3	23.4	16.9	22.2
LLaVA-OneVision-7B	55.4	57.0	**35.2**	49.2
Qwen2.5-VL-3B	47.7	34.2	9.6	30.5
Qwen2.5-VL-7B	62.2	39.4	17.6	39.7
Qwen3-VL-4B	73.0	64.0	16.6	51.2
Qwen3-VL-8B	74.5	**68.3**	28.5	**57.1**
Domain-Specific Remote Sensing MLLMs				
GeoChat	25.7	26.5	31.6	27.9
EarthDial	39.8	18.2	21.9	26.6
VHM	29.4	22.9	16.3	22.9
GeoLLaVA-8K	28.0	33.6	22.3	28.0
ZoomEarth	44.0	39.9	30.8	38.2
Geoeyes	55.1	50.6	24.8	43.5
GeoVista	65.2	61.1	23.0	49.8
Self-Evolving				
TerraEvo (Ours)	**75.5**	67.1	25.1	55.9

Table 9: RSHR-Bench per-category accuracy (%) across Perception, Reasoning, and Multi-turn tasks. Perception: COL = Color, SHP = Shape, DET = Detection, OC = Object Classification, REL = Object Spatial Relationship, OGD = Object Grounding, RG = Regional Grounding, OCN = Object Counting, RCN = Regional Counting. Reasoning: AD = Anomaly Detection, FP = Future Prediction, MRJC = Multi-region Joint Contrast, MRJCS = Multi-region Joint Contrast (single-image), OSJ = Object State Judgment. Multi-turn: MAD = Anomaly, MTFP = Future Prediction, MOSJ = Object State Judgment. Avg. denotes the unweighted mean within each group; Final Avg. denotes the unweighted mean of the three group averages. Bold indicates the best result in each column, including ties.

	Perception										Reasoning						Multi-turn				
Model	**COL**	**SHP**	**DET**	**OC**	**REL**	**OGD**	**RG**	**OCN**	**RCN**	**Avg.**	**AD**	**FP**	**MRJC**	**MRJCS**	**OSJ**	**Avg.**	**MAD**	**MTFP**	**MOSJ**	**Avg.**	**Final Avg.**
Closed-Source MLLMs																					
GPT-4o (Hurst et al., 2024)	46.5	21.0	32.0	30.5	39.5	31.0	28.6	22.0	34.0	31.7	66.0	48.0	50.0	42.0	72.0	55.6	75.0	66.7	81.1	74.3	53.9
GPT-5 (Singh et al., 2025)	54.5	19.0	31.0	33.5	32.5	37.5	41.4	20.5	43.0	34.8	78.0	34.0	50.0	40.0	66.0	53.6	70.0	71.3	84.1	75.1	54.5
Open-Source MLLMs																					
InternVL2.5-8B (Chen et al., 2024a)	58.5	21.0	21.0	21.5	49.0	30.5	21.4	17.0	41.0	31.2	66.0	54.0	60.0	46.0	58.0	56.8	80.0	67.3	76.5	74.6	54.2
InternVL3.5-8B (Wang et al., 2025b)	48.0	21.0	36.0	43.0	21.0	34.5	35.7	28.0	36.0	33.7	62.0	54.0	40.0	34.0	62.0	50.4	75.0	65.3	77.3	72.5	52.2
InternVL3-8B (Zhu et al., 2025)	48.0	22.0	26.0	53.0	39.0	20.0	27.1	20.5	30.0	31.7	64.0	44.0	30.0	30.0	50.0	43.6	66.7	68.7	80.3	71.9	49.1
LLaVA1.5-7B (Liu et al., 2024)	25.0	24.0	25.0	26.0	25.0	25.0	22.9	25.0	25.0	24.8	24.0	24.0	35.0	34.0	22.0	27.8	25.0	24.7	48.5	32.7	28.4
LLaVA-OneVision-7B (Li et al., 2024)	49.5	24.0	26.0	29.5	46.5	26.0	24.3	21.0	35.0	31.3	56.0	42.0	45.0	44.0	52.0	47.8	65.0	59.3	70.5	64.9	48.0
Qwen2.5-VL-3B (Bai et al., 2025b)	41.5	22.0	26.0	27.0	46.5	25.0	25.7	21.0	38.0	30.3	66.0	50.0	55.0	32.0	44.0	49.4	71.7	64.0	51.5	62.4	47.4
Qwen2.5-VL-7B (Bai et al., 2025b)	46.5	22.0	30.0	36.0	50.0	26.5	28.6	20.0	30.0	32.2	70.0	32.0	55.0	36.0	58.0	50.2	65.0	65.3	69.7	66.7	49.7
Qwen3-VL-4B (Bai et al., 2025a)	49.5	27.0	21.0	34.5	48.5	35.0	44.3	19.0	41.0	35.5	78.0	40.0	50.0	40.0	66.0	54.8	68.3	64.7	84.1	72.4	54.2
Qwen3-VL-8B (Bai et al., 2025a)	53.5	23.0	24.0	33.5	45.0	44.0	48.6	18.5	36.0	36.2	76.0	48.0	50.0	32.0	62.0	53.6	66.7	66.7	81.1	71.5	53.8
Domain-Specific Remote Sensing MLLMs																					
GeoChat (Kuckreja et al., 2024)	25.0	24.0	25.0	25.0	25.0	25.0	22.9	25.0	25.0	24.7	28.0	24.0	35.0	34.0	26.0	29.4	28.3	26.7	49.2	34.7	29.6
EarthDial (Soni et al., 2025)	40.5	30.0	23.0	37.5	31.5	32.5	28.6	19.0	35.0	30.8	48.0	46.0	30.0	40.0	58.0	44.4	55.0	53.3	78.0	62.1	45.8
VHM (Pang et al., 2025)	25.0	24.0	24.0	25.0	36.5	25.0	22.9	21.5	25.0	25.4	24.0	24.0	50.0	34.0	20.0	30.4	23.3	26.7	47.7	32.6	29.5
GeoLLaVA-8K (Wang et al., 2026a)	42.5	21.0	28.0	26.5	6.5	27.5	24.3	35.0	37.0	27.6	44.0	48.0	55.0	36.0	54.0	47.4	48.3	63.3	78.0	63.2	46.1
ZoomEarth (Liu et al., 2026b)	48.5	20.0	22.0	19.5	33.0	25.0	21.4	26.5	33.0	27.7	44.0	30.0	55.0	26.0	52.0	41.4	55.0	58.0	61.4	58.1	42.4
Geoeyes (Wang et al., 2026b)	79.5	49.0	34.0	36.0	54.5	42.5	45.7	22.0	53.0	46.2	60.0	40.0	65.0	28.0	64.0	51.4	60.0	64.7	77.3	67.3	55.0
GeoVista (Zhu et al., 2026)	57.0	37.0	34.0	48.0	36.5	28.5	21.4	30.0	30.0	35.8	64.0	50.0	60.0	28.0	66.0	53.6	65.0	63.3	70.5	66.3	51.9
Self-Evolving																					
TerraEvo (Ours)	46.5	27.0	28.0	38.5	43.5	35.0	42.9	22.5	43.0	36.3	76.0	34.0	55.0	42.0	64.0	54.2	71.7	68.0	83.3	74.3	**55.0**

Table 10: Experimental results on the perception and reasoning dimensions on XLRS-Bench, including our self-evolving framework. The sub-tasks include Perception: Overall/Regional Counting (OC/RC), Overall/Regional Land Use Classification (OLUC/RLUC), Object Classification (OCC), Object Color (OCL), Object Motion State (OMS), and Object Spatial Relationship (OSR); and Reasoning: Anomaly Detection and Interpretation (AD), Environmental Condition Reasoning (ECR), Route Planning (RP), Regional Counting with Change Detection (RCCD), and Counting with Complex Reasoning (CCR). 'Avg.' represents the macro average accuracy across sub-tasks. Bold indicates the best result in each column, including ties.

Method	Perception								Reasoning					Avg.
Sub-tasks (L-3 Capability)	**OC**	**RC**	**OLUC**	**RLUC**	**OCC**	**OCL**	**OMS**	**OSR**	**AD**	**ECR**	**RP**	**RCCD**	**CCR**	
Closed-Source MLLMs														
GPT-4o (Hurst et al., 2024)	21.7	35.0	22.0	67.5	41.5	27.0	66.7	32.0	66.0	78.0	43.0	38.3	40.0	44.5
GPT-5 (Singh et al., 2025)	26.7	0.0	26.0	72.0	49.0	38.4	65.0	35.6	69.0	**83.0**	35.0	21.7	40.0	43.2
Open-Source MLLMs														
InternVL2.5-8B (Chen et al., 2024a)	36.7	39.0	9.0	77.0	33.0	35.2	66.7	20.8	73.0	82.0	37.0	55.0	39.0	46.4
InternVL3.5-8B (Wang et al., 2025b)	28.3	42.0	47.0	75.0	33.4	26.1	65.0	**42.4**	68.0	77.0	52.0	**56.7**	50.0	51.0
InternVL3-8B (Zhu et al., 2025)	33.3	37.0	15.0	66.0	42.8	31.2	65.0	27.2	75.0	77.0	32.0	33.3	50.0	45.0
LLaVA1.5-7B (Liu et al., 2024)	16.7	29.0	0.0	24.5	21.0	16.8	35.0	23.8	35.0	48.0	46.0	28.3	21.0	26.5
LLaVA-OneVision-7B (Li et al., 2024)	31.7	40.0	2.0	62.5	39.8	33.9	65.0	27.0	**79.0**	81.0	29.0	48.3	32.0	43.9
Qwen2.5-VL-3B (Bai et al., 2025b)	28.3	41.0	4.0	72.5	34.2	33.4	65.0	29.4	70.0	74.0	19.0	33.3	43.0	42.1
Qwen2.5-VL-7B (Bai et al., 2025b)	25.0	40.0	29.0	75.0	37.2	39.1	66.7	33.8	65.0	73.0	25.0	40.0	46.0	45.8
Qwen3-VL-4B (Bai et al., 2025a)	31.7	46.0	21.0	79.0	41.8	42.8	65.0	28.4	75.0	76.0	43.0	46.7	49.0	49.6
Qwen3-VL-8B (Bai et al., 2025a)	20.0	**48.0**	26.0	**82.0**	46.4	43.1	66.7	30.4	75.0	79.0	37.0	46.7	51.0	50.1
Domain-Specific Remote Sensing MLLMs														
GeoChat (Kuckreja et al., 2024)	16.7	29.0	0.0	23.0	21.1	16.8	35.0	24.4	36.0	48.0	46.0	28.3	21.0	26.6
EarthDial (Soni et al., 2025)	21.7	41.0	2.0	39.5	33.0	29.9	66.7	24.0	66.0	69.0	31.0	46.7	46.0	39.7
VHM (Pang et al., 2025)	16.7	30.0	9.0	23.0	10.6	16.8	35.0	25.8	34.0	48.0	45.0	28.3	21.0	26.4
GeoLLaVA-8K (Wang et al., 2026a)	26.7	38.0	**48.0**	69.0	41.8	31.5	65.0	35.4	67.0	78.0	**66.0**	50.0	**52.0**	51.4
ZoomEarth (Liu et al., 2026b)	21.7	31.0	9.0	36.0	29.4	30.8	38.3	16.8	55.0	56.0	37.0	18.3	23.0	30.9
Geoeyes (Wang et al., 2026b)	**38.3**	40.0	24.0	73.5	**59.5**	**66.1**	68.3	32.2	77.0	80.0	56.0	40.0	50.0	**54.2**
GeoVista (Zhu et al., 2026)	33.3	41.0	23.0	77.0	50.8	45.4	**70.0**	36.6	56.0	80.0	54.0	38.3	44.0	50.0
Self-Evolving														
TerraEvo (Ours)	28.3	**48.0**	30.0	78.0	45.2	43.1	66.7	28.6	76.0	76.0	36.0	51.7	48.0	50.4