GEOREFLEX: SELF-CORRECTING ACTIVE PERCEPTION FOR ULTRA-HIGH-RESOLUTION REMOTE SENSING

Anonymous authors
Paper under double-blind review

ABSTRACT

Ultra-high-resolution (UHR) remote-sensing understanding requires integrating global context and local detail within a limited visual budget. Active-perception methods acquire evidence through dynamic cropping, but incorrect crops can leave subsequent exploration and reasoning dependent on incomplete observations. We introduce **GeoReflex**, an active-perception framework that uses reasoning to examine and correct multi-step visual observations. It connects evidence diagnosis with observation adjustment: when evidence is insufficient, the model can restore the parent view and use its diagnosis to redirect exploration. To learn this behavior, we construct **GeoReflex-48K** with $47{,}655$ examples and combine the capabilities of a Generator, Reflector, and Refiner in a single model through supervised fine-tuning and reflection-efficient reinforcement learning, teaching when to reflect and how to correct. Across three remote-sensing benchmarks, GeoReflex exceeds the strongest compared external baselines on RSHR-Bench and XLRS-Bench by 5.40 and 7.04 points, respectively, demonstrating the value of recovery-oriented post-training for UHR understanding. Our anonymous repository is available at https://anonymous.4open.science/r/GeoReflex-CB14/.

1 INTRODUCTION

UHR remote-sensing questions can depend on small objects, distant regions, and global–local relations distributed across a large visual field. Under a limited visual-token budget, a model must therefore *collect sufficient task-relevant evidence* before committing to an answer. Yet more observations are not inherently better: premature stopping omits clues, whereas unfocused or repeated zooming consumes context without adding evidence. Reasoning from partial observations can then rationalize an answer unsupported by the image. UHR understanding thus requires deciding both *where to look* and *when the evidence is sufficient*.

Recent active-perception methods, including ZoomEarth (Liu et al., 2026b), GeoEyes (Wang et al., 2026b), GeoLens (Wang et al., 2026c), and GeoVista (Zhu et al., 2026), improve task-conditioned evidence acquisition through multi-scale observation, visual tools, or multi-branch exploration. These advances motivate a complementary learning problem: how to recognize incomplete evidence after an incorrect crop or overlooked region, and revise the observation path. Successful exploration alone does not provide explicit supervision for recovering from such failures.

This limitation suggests the reverse question: *can reasoning improve perception?* Reasoning represents the evidence required by the task, such as both a target and a reference landmark, a connected route, or complete coverage for counting. Reflection can compare those requirements with acquired evidence and convert missing or inconsistent information into a targeted correction. Figure 1 illustrates this feedback: an attribute mismatch in the current observation redirects perception.

Preliminary study: does tool access suffice? Before introducing additional training, we test whether recovery can be elicited by giving Qwen3-VL-8B-Thinking a parent-view restoration tool and its usage instructions (Fig. 1, lower left), with only nine fallback calls across 8,750 questions. Merely providing recovery tools does not consistently improve performance, motivating explicit supervision for evidence diagnosis and corrective observation.

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.

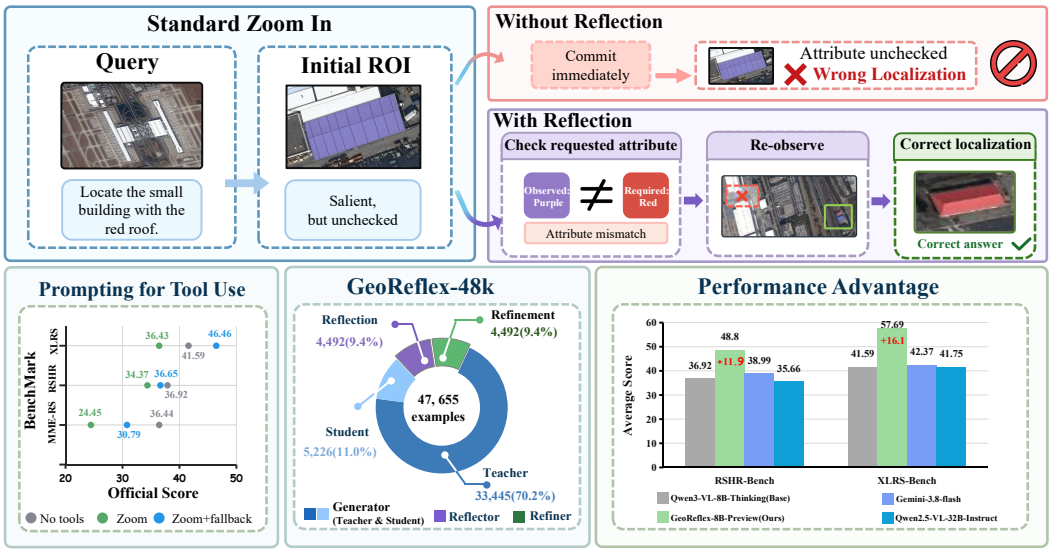


Figure 1: **GeoReflex overview.** Top: attribute checking redirects an incorrect crop. Bottom: tool-prompting controls without additional training, GeoReflex-48K supervision, and post-training performance. No-tool scores are a main-evaluation reference; performance gains are relative to Qwen3-VL-8B-Thinking.

Motivated by this gap, we introduce **GeoReflex**, a post-training framework that teaches a small vision-language model to diagnose insufficient visual evidence and revise its observation path. A single inference-time policy assumes three responsibilities: a **Generator** actively inspects and answers, a **Reflector** audits the supporting evidence and can invoke recovery, and a **Refiner** continues from the restored state. Recovery changes the visual context rather than merely appending a critique: it restores the parent view and exposes the rejected path for the next decision. Our contributions are:

1. We frame UHR remote-sensing understanding as a visual perception problem that can be improved through reasoning, using reflection to diagnose insufficient evidence and guide corrective observation.
2. We construct **GeoReflex-48K**, a role-aligned remote-sensing recovery corpus with 47,655 examples: successful trajectories supervise generation, failed prefixes paired with teacher diagnoses supervise reflection, and restored states supervise the next corrective decision.
3. We post-train Qwen3-VL-8B-Thinking with SFT and **reflection-efficient GRPO**, evaluating learned correction against active-perception baselines on three RS benchmarks.

2 Related Work

Active perception for UHR remote sensing. General-purpose MLLMs (Achiam et al., 2023; Team et al., 2023; Liu et al., 2023; Bai et al., 2023; Wang et al., 2024; Bai et al., 2025) and remote-sensing VLMs (Hu et al., 2025; Zhan et al., 2025; Kuckreja et al., 2024; Shu et al., 2025; Wang et al., 2026d) provide the foundations for visual agents, but UHR imagery requires balancing spatial coverage and fine-grained detail under a limited visual-token budget. Higher static resolution, including inputs at roughly 8K (Wang et al., 2026a), increases visible detail but retains the coverage–detail trade-off under a fixed context budget. Textual reasoning (Liu et al., 2026a; Fiaz et al., 2025; Wang et al., 2025; Zhang et al., 2026) structures inference over available evidence without itself acquiring missing observations. Active perception instead obtains task-conditioned views during inference. In natural images, interleaved visual tool use demonstrates that cropping and retrieval can be learned (Su et al., 2025; Zheng et al., 2026; Xien Chng et al., 2025; Huang et al., 2026). In remote sensing, crop selection controls the spatial extent and fine-grained detail presented to the model without changing the source image's physical ground sample distance. For remote sensing, ZoomEarth (Liu et al., 2026b), GeoEyes (Wang et al., 2026b), and GeoLens (Wang et al., 2026c) use multi-scale or multi-tool observation, supported by interleaved inspection trajectories in LRS-

GRO, UHR-CoZ, and GeoMTVR. Zoom-and-search methods (Zhou et al., 2025; Zhang et al., 2024) select relevant regions through sequential inspection or retrieval, while GeoVista (Zhu et al., 2026) supports planning-driven, multi-branch exploration through Observe–Plan–Track. Successful exploration trajectories teach evidence acquisition; GeoReflex complements this supervision by pairing executed visual failures with diagnoses of insufficient evidence and corrective continuations that revise the observation path.

From reflection to self-correction. Self-correction typically alternates generation, evaluation, and revision, with its effectiveness depending on feedback reliability (Kamoi et al., 2024). Intrinsic approaches use model-generated critiques (Yao et al., 2022; Shinn et al., 2023; Madaan et al., 2023; Bai et al., 2022; Dhuliawala et al., 2024), but self-evaluation can be overconfident and fail to improve initial predictions (Huang et al., 2024; Kamoi et al., 2024; Renze & Guven, 2024). Externally grounded methods instead use tests, interpreters, or retrieved evidence to identify errors and guide revision (Gao et al., 2023; Chen et al., 2024; Gou et al., 2024); their applicability depends on access to suitable feedback at inference. Training-time revision can teach models to correct their outputs when supervision is available (Welleck et al., 2022; Xi et al., 2024; Xie et al., 2025; Kumar et al., 2025), but this line of work has been developed mainly for language, mathematics, and code, not for recovering from an already executed remote-sensing observation. The resulting gap is both behavioral and methodological: a completed, typically greedy decoding trajectory can commit to an incorrect visual path, and the agent has no external signal for whether that commitment should be retained. GeoReflex decomposes this setting into Generator, Reflector, and Refiner responsibilities. An offline teacher labels reflection and refinement trajectories, while paired supervision distills both into the same student policy: it invokes fallback after an executed failure and continues from the restored parent view without an online critic model.

3 METHOD

GeoReflex trains a single policy to improve UHR perception through reasoning about missing evidence and recovery from incorrect observations. A plausible crop can omit a required object or spatial relation. The key idea is to make diagnosis actionable: return to the parent view, retain the failed-path context, and choose a better observation.

Figure 2 summarizes two-stage post-training. SFT combines exploration, diagnosis, and corrective continuation in one model; reflection-efficient GRPO learns when correction is warranted from interaction outcomes and training-only feedback.

Problem formulation. The input is a task $x = (I, q)$ comprising a UHR image I and a question q; the output is a final answer $\hat{a}$, whose form depends on the task. A policy π_θ with parameters θ generates responses y_i at interaction step i, each parsed into an action a_i: inspect a crop, request recovery, or answer. The environment returns observation o_i and maintains the state

$$h_i = (m_i, v_i), \tag{1}$$

where m_i is the message and observation history and v_i is the current view, with parent relations retained by the environment. The resulting interaction is

$$\tau = (x, h_1, y_1, a_1, o_1, \ldots, h_T, y_T, a_T, o_T), \qquad \tau \sim \Pi_\theta(\cdot \mid x), \tag{2}$$

where T is the number of model calls and Π_θ is the rollout distribution induced by the policy and environment. Each response is conditioned on the updated visual state. Appendix A specifies input construction and action execution.

The learning objective is to maximize expected trajectory reward under bounded interaction:

$$\max_\theta \ \mathbb{E}_{x\sim\mathcal{D}_{\text{task}},\, \tau\sim\Pi_\theta(\cdot|x)}[R(\tau)], \qquad T \le B_{\text{call}}, \quad N_{\text{fb}}(\tau) \le B_{\text{fb}}. \tag{3}$$

Here $\mathcal{D}_{\text{task}}$ is the training-task distribution, R scores answer quality and recovery (Eq. 8), and N_{fb} counts accepted recoveries. The limits B_{call} and B_{fb} are eight calls and three recoveries in main evaluation; training limits appear in Appendix E.

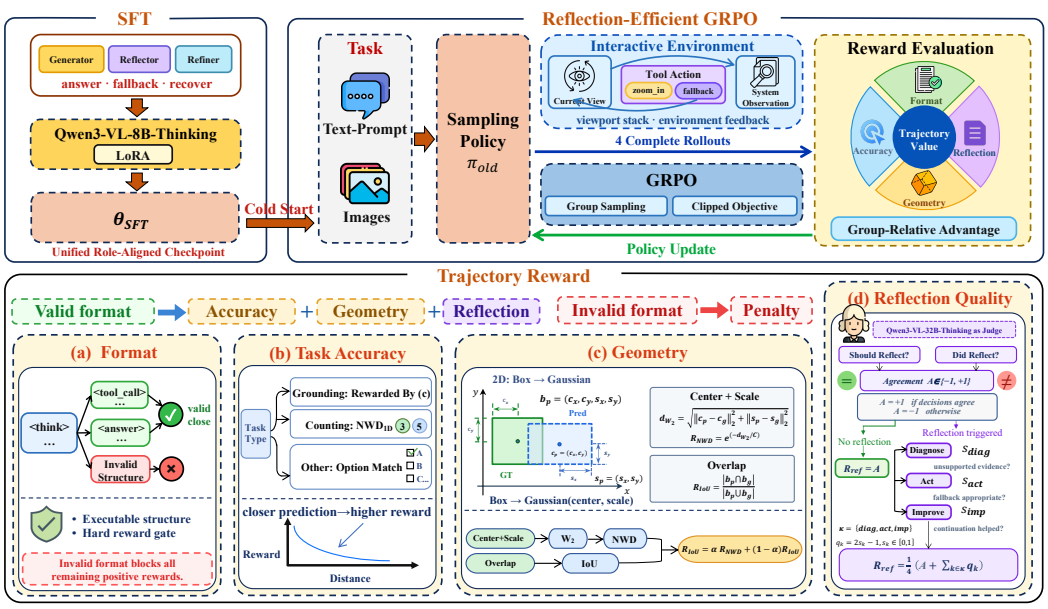


Figure 2: **Two-stage post-training.** SFT integrates Generator, Reflector, and Refiner capabilities into one policy; GRPO optimizes complete rollouts using format-gated accuracy, geometry, and reflection rewards. The 32B judge is training-only.

3.1 Diagnosis-Guided Visual Recovery

From diagnosis to visual correction. Crop acquisition exposes local detail but can also remove the context needed to judge whether that detail answers the question. GeoReflex makes this judgment actionable: a diagnosis identifies what the current view cannot support, parent-view restoration reopens the spatial context, and corrective exploration seeks the missing evidence. Reasoning thus guides what the agent observes next. Let b_i denote the rejected region's bounding box, c_i the diagnosed cause, and $u_i \in \{\textsc{Replan}, \textsc{Answer}\}$ the model's intended next step. An accepted recovery action implements this transition by restoring the parent view and retaining the failed-path note:

$$v_{i+1} = \mathrm{parent}(v_i), \qquad o_i^{\mathrm{fb}} = \mathcal{O}_{\mathrm{fb}}(v_{i+1}, v_i, b_i, c_i, u_i), \qquad m_{i+1} = m_i \oplus (y_i, o_i^{\mathrm{fb}}). \tag{4}$$

Here $\mathcal{O}_{\mathrm{fb}}$ constructs the recovery observation and $\oplus$ appends it with the model response to the history. Diagnosis identifies insufficient evidence; fallback changes the active view. Textual reflection need not invoke fallback. Restoration changes the observation state without erasing the interaction history, allowing subsequent reasoning to account for why the previous crop was rejected. The failed region remains revisitable if later evidence supports it. Boundary conditions are specified in Appendix A.

Focused evidence acquisition. Reflection efficiency requires selecting useful corrections. The reflection reward penalizes unnecessary reflection, rewards justified non-reflection, and evaluates diagnosis, action appropriateness, and expected improvement when reflection occurs (Eq. 11).

Generator–Reflector–Refiner. Three capabilities share one policy: the Generator acquires evidence and proposes an answer, the Reflector diagnoses omissions or inconsistencies and may request recovery, and the Refiner continues from the resulting context. These roles can interleave within a rollout; they are neither separate networks nor mandatory successive calls. For example, a candidate object may match the requested appearance but violate a spatial relation. The diagnosis identifies that mismatch, while restoring the parent view allows the next observation to include the missing reference landmark. Retaining the failed-path context connects the new observation to the reason for reconsideration.

Learning recovery requires supervision where an observation path needs revision. Successful trajectories show how evidence supports an answer but leave the diagnosis and continuation after failure unspecified. We therefore pair failed prefixes with diagnoses and next decisions from restored views.

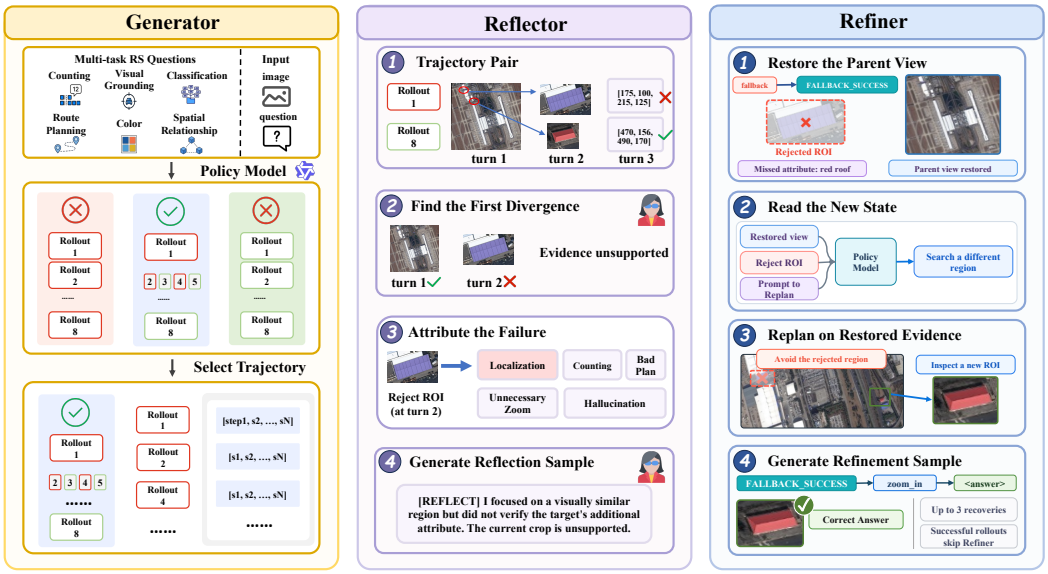


Figure 3: **Recovery-data construction.** Sampled failures are diagnosed and paired with next-turn targets from restored parent views, linking reflection to corrective action.

GeoReflex-48K translates this framework into supervision that teaches all three capabilities without an external critic at inference.

3.2 GeoReflex-48K: Learning Exploration, Reflection, and Refinement

We construct **GeoReflex-48K** with 47,655 examples to teach exploration, reflection, and refinement within one model. APE-GRO demonstrations from GeoVista (Zhu et al., 2026) and successful student trajectories teach evidence acquisition; paired diagnoses and continuations teach recovery (Fig. 3). The corpus covers classification, counting, spatial relationships, route planning, object color, and visual grounding.

Generating attempts and diagnosing failures. We sample multiple student attempts per question, retaining failures from both mixed-outcome questions and questions with no successful attempt. Qwen3-VL-32B-Thinking diagnoses each failure and writes a recovery turn, using a successful sibling as reference when available. Sampling and filtering details appear in Appendix C.

Pairing reflection with corrected continuation. Each teacher rewrite yields two examples implementing Eq. 4. The reflection example preserves the failed prefix and supervises the diagnosis and recovery decision. The refinement example appends the restored observation and supervises the single next turn. This pairing connects recognizing an error with acting on it. Both examples retain the same executed actions and observations, so the corrective decision is learned in the context that produced the failure. The refinement target demonstrates how to resume from that state, rather than replacing the failed attempt with an unrelated successful solution.

SFT supervises all retained assistant turns, including the failure prefix, rather than only the new target. Appendix E specifies loss masking; Appendix A describes coordinate and image handling.

Pair-level filtering and corpus composition. We discard both examples if either fails transition, image, coordinate, or answer checks. After deduplication, 4,492 pairs cover 1,034 questions. Combined with 33,445 teacher demonstrations and 5,226 student successes, reflection and refinement each account for 9.4% of the corpus (Fig. 1). Successful examples include both active exploration and direct answers.

3.3 TWO-STAGE POST-TRAINING

Stage I: Supervised Fine-Tuning. A recovery example is useful only if the same policy can both recognize the failure and continue from the corrected context. SFT therefore learns ordinary exploration and recovery jointly. Let $\mathcal{D}_G$ contain teacher and successful-student trajectories, and let $\mathcal{D}_R$ and $\mathcal{D}_F$ contain the paired reflection and refinement examples constructed above. The training corpus is

$$\mathcal{D}_{\mathrm{SFT}} = \mathcal{D}_G \cup \mathcal{D}_R \cup \mathcal{D}_F. \tag{5}$$

For example d, let $c_{d,t}$ denote the complete multimodal context available before token $y_{d,t}$, including preceding model messages and environment observations. Let $\mu_{d,t}$ indicate whether that token is a supervised policy output. The masked token objective is

$$\mathcal{L}_{\mathrm{SFT}}(\theta) = -\sum_{d \in \mathcal{D}_{\mathrm{SFT}}} \sum_{t} \mu_{d,t} \log \pi_\theta(y_{d,t} \mid c_{d,t}). \tag{6}$$

The mask excludes task inputs and environment observations. Shared parameters connect answering with recovery; the full multimodal context includes intervening observations.

We fine-tune Qwen3-VL-8B-Thinking with LoRA on the complete corpus. Appendix E provides the training configuration and completed-run statistics.

Stage II: Reflection-Efficient GRPO. Demonstrations specify how to recover, but the policy must also learn when recovery is useful under its own observations. We therefore initialize reinforcement learning from θ_{SFT} and optimize complete interactions with Group Relative Policy Optimization (GRPO) (Shao et al., 2024). For each task, the sampling policy $\pi_{\theta_{\mathrm{old}}}$ produces a group of K rollouts:

$$\mathcal{G}_x = \{\tau_k\}_{k=1}^{K}, \qquad \tau_k \sim \Pi_{\theta_{\mathrm{old}}}(\cdot \mid x). \tag{7}$$

Rewarding correct answers and warranted recovery. Answer correctness alone does not distinguish a necessary correction from an unnecessary detour that reaches the same answer. Conversely, a useful recovery may improve the evidence without yet producing a correct final response. We therefore assess both task outcomes and reflection behavior: accuracy and geometry reward the answer, while reflection supervision evaluates the decision to reconsider and the quality of the ensuing correction.

Each rollout is scored after its model calls and environment transitions are complete. The reward combines task success with the quality of the recovery process, while gating on executable output format:

$$R(\tau) = \begin{cases} \omega_{\mathrm{fmt}} r_{\mathrm{fmt}} + \omega_{\mathrm{acc}} r_{\mathrm{acc}} + \omega_{\mathrm{geo}} r_{\mathrm{geo}} + \omega_{\mathrm{ref}} r_{\mathrm{ref}}, & r_{\mathrm{fmt}} = 1, \\ -\gamma_{\mathrm{fmt}}, & r_{\mathrm{fmt}} = 0. \end{cases} \tag{8}$$

Here $r_{\mathrm{fmt}} \in \{0, 1\}$ indicates format validity, ω denotes the component weights, and $\gamma_{\mathrm{fmt}} > 0$ penalizes invalid output. The hard gate prevents malformed or non-executable trajectories from receiving the remaining rewards.

Accuracy and geometry. For a predicted box $B = (c_x, c_y, w, h)$ and target B^*, we use the Gaussian representation $\mathcal{N}(c, \Sigma_B)$ with $c = (c_x, c_y)$ and $\Sigma_B = \mathrm{diag}(w^2/4, h^2/4)$. Following normalized Wasserstein distance (NWD) (Wang et al., 2021), the squared 2-Wasserstein distance between the box representations reduces to

$$\begin{aligned} W_2^2(B, B^*) &= \|c - c^*\|_2^2 + \left\| \tfrac{1}{2}(w, h) - \tfrac{1}{2}(w^*, h^*) \right\|_2^2, \\ \mathrm{NWD}(B, B^*) &= \exp\left(-\frac{\sqrt{W_2^2(B, B^*)}}{C_B} \right). \end{aligned} \tag{9}$$

Here $C_B > 0$ controls the distance scale in the same coordinate units as the box parameters. NWD remains informative when two small boxes do not overlap, while IoU directly measures shared area. We combine these complementary spatial signals:

$$r_{\mathrm{geo}} = \lambda_{\mathrm{nwd}} \, \mathrm{NWD}(B, B^*) + (1 - \lambda_{\mathrm{nwd}}) \, \mathrm{IoU}(B, B^*). \tag{10}$$

Here $\lambda_{\mathrm{nwd}} \in [0, 1]$. Grounding accuracy uses NWD, so accuracy and geometry share that signal; geometry additionally rewards overlap. Counting uses one-dimensional NWD and multiple-choice accuracy uses exact match. Geometry is zero when boxes are unavailable. Appendix E details counting, multi-box aggregation, and the combined grounding reward.

Reflection behavior. A training-only Qwen3-VL-32B-Thinking judge assesses whether reflection is warranted, $z^\star \in \{0, 1\}$. Detected reflection $z \in \{0, 1\}$ records a fallback call or reflection marker, not recovery success. Their agreement is $A = 2\mathbf{1}[z = z^\star] - 1$. For reflected trajectories, raw scores $s_j \in [0, 1]$ assess the identified mistake, fallback appropriateness, and expected progress toward the reference answer, using the recovery turn when available. For $j \in \{\mathrm{diag}, \mathrm{act}, \mathrm{imp}\}$, we map $q_j = 2s_j - 1$ and compute

$$r_{\mathrm{ref}} = \frac{A + z\,(q_{\mathrm{diag}} + q_{\mathrm{act}} + q_{\mathrm{imp}})}{1 + 3z}. \tag{11}$$

When $z = 0$, only agreement contributes; otherwise, all four assessments are averaged. The denominator counts active assessments, keeping both cases in $[-1, 1]$. The judge assesses whether an error is recoverable, rather than simply whether the answer is wrong. It marks high- or medium-recoverability errors as warranting reflection; correct trajectories and purely fine-grained recognition errors do not warrant it under this taxonomy. Thus, an incorrect answer alone does not automatically make reflection desirable. This rewards justified non-reflection and useful correction without a length penalty. Judge criteria and prompts appear in Appendix B.

Group-relative optimization. We normalize rewards within each rollout group and optimize the dual-clipped, token-mean GRPO objective over policy outputs, excluding task inputs and tool observations. Appendix F gives the objective.

4 Experiments

We first evaluate the complete policy on three benchmarks, then isolate reflection-reward contributions during training and diagnosis-guided recovery at inference. Finally, we examine the complete policy's performance and resource use under different interaction budgets.

4.1 Setup

Models. We evaluate two checkpoints derived from Qwen3-VL-8B-Thinking: GeoReflex-8B-Instruct after SFT and GeoReflex-8B-Preview after GRPO. Table 1 compares both with model and active-perception baselines. Baselines are cited in Sec. 2. Qwen3-VL-32B-Thinking constructs offline targets and judges GRPO trajectories but is absent at test time.

Benchmarks and metrics. We report results on RSHR-Bench, XLRS-Bench, and MME-RealWorld-RS using each benchmark's official evaluation functions for all tasks and its official score aggregation. For RSHR-Bench, we report the overall score. Matched-prefix recovery scores are reported separately from full-benchmark results.

Inference protocol and visual preprocessing. We evaluate all reported models ourselves except for the GeoLens RSHR and XLRS results, which are taken from Wang et al. (2026c). Our main evaluations cap interactive methods at eight turns and otherwise retain each method's official tools and evaluation settings. All benchmark scores use the official benchmark evaluation programs. GeoReflex uses the training-time `zoom_in` and `fallback` protocol with a normalized $[0, 1000]$ viewport stack and at most three recoveries; the same wrapper serves its SFT and GRPO checkpoints. Appendix A details preprocessing and the scope of protocol comparability.

4.2 Main results

GeoReflex-8B-Preview achieves the highest RSHR and XLRS scores among the methods in Table 1, exceeding the strongest external baseline on each by 5.40 and 7.04 points, respectively. It improves over the SFT checkpoint on all three benchmarks, while remaining below Gemini-3.8-flash on MME-RS. The corresponding gains over SFT are 7.54, 7.79, and 8.61 points on RSHR, XLRS, and MME-RS, respectively. The gains across all three benchmarks show that reinforcement learning improves the SFT-initialized policy beyond demonstration fitting. They support recovery-oriented post-training as an effective approach to UHR understanding; the following controls examine the contributions of reward design and executable recovery.

4.3 Reward Ablation and Mechanism Analysis

Reward components. To isolate reward contributions, all variants share the SFT initialization, crop and fallback tools, training data, rollout budgets, update counts, and evaluation settings. Each reduced variant sets one reward weight to zero; all other weights remain fixed.

Format and accuracy remain enabled while geometry or reflection reward is removed. Removing geometry retains the grounding NWD accuracy signal but removes the additional geometric reward. The variant without reflection reward retains reflection-oriented SFT supervision and access to fallback, isolating the incremental contribution of reflection reward rather than the entire recovery framework.

Removing geometry lowers RSHR, XLRS, and MME-RS by 2.79, 5.22, and 6.66 points, respectively, indicating the importance of geometric supervision. This intervention removes the IoU contribution and reduces the effective NWD weight, while retaining NWD in the accuracy term; it does not isolate IoU alone. Removing reflection reward lowers the same scores by 0.85, 1.74, and 3.42 points. Because both variants retain fallback access, this comparison supports learning how to use recovery through explicit behavioral feedback, rather than merely providing a recovery tool. Together, the ablations support the full objective's combination of spatially grounded task supervision and reflection assessment.

Table 1: Complete three-benchmark comparison. RSHR denotes the overall score; Appendix D reports task breakdowns.

Model	RSHR	XLRS	MME-RS
Closed-source MLLMs			
GPT-4o	32.95	32.15	41.93
Gemini-3.8-flash	38.99	42.37	**59.07**
Claude-4.5-Sonnet	32.77	37.63	45.80
Open-source MLLMs			
DeepSeek-VL2-small	35.22	40.32	27.74
LLaVA1.5-7B	29.19	35.10	22.12
LLaVA-UHD-v3	35.85	36.53	45.37
LLaVA-OneVision-2	35.41	38.54	<u>57.12</u>
Large-Format MLLMs (2k+ level)			
Qwen3-VL-8B-Instruct	38.80	43.12	46.33
Qwen3-VL-8B-Thinking	36.92	41.59	36.44
GLM-4.6V-Flash	40.63	43.38	48.98
InternVL3.5-8B	36.85	41.72	50.67
Qwen2.5-VL-32B-Instruct	35.66	41.75	23.22
Remote Sensing MLLMs			
GeoLLaVA-8K	31.39	37.11	23.43
GeoChat	29.37	26.72	28.12
VHM	27.10	29.51	22.95
EarthDial	30.88	33.34	28.89
Methods			
ZoomEarth	31.20	34.81	43.15
ZoomSearch	31.51	40.68	44.46
ImageRag	34.72	40.19	44.68
Coarse-to-fine	26.79	35.65	41.36
GeoEyes	39.75	42.34	22.34
DeepEyes	<u>43.40</u>	35.97	40.50
RS-EoT	31.14	37.14	31.43
RSCoVLM	33.39	42.95	28.14
GeoLens	40.50	49.12	48.42
GeoVista-7B-Preview	41.38	<u>50.65</u>	40.05
Ours			
GeoReflex-8B-Instruct	41.26	49.90	42.43
GeoReflex-8B-Preview	**48.80**	**57.69**	51.04

From reflection to refinement. We examine how the Reflector's diagnostic context informs the Refiner's continuation by independently varying diagnosis inclusion and parent-view restoration (Table 3). The cohort contains 82 eligible questions from 3,080 XLRS questions, with one trajectory per question, and covers 2.66% of the full evaluation; this fraction is not an estimate of overall reflection incidence. All interventions share the selected reflection-round prefix, even for trajectories with multiple reflection events; both diagnosis conditions reuse the same text. These controls probe the reflection-to-refinement mechanism within one trained policy, not the removal of separately trained role components.

Table 2: Reward-group ablation with matched training and tools. Checkmarks indicate active components; the full objective performs best on all three benchmarks.

	Reward configuration				Benchmark score		
Variant	Format	Accuracy	Geometry	Reflection	RSHR	XLRS	MME-RS
Full objective	✓	✓	✓	✓	**48.80**	**57.69**	**51.04**
Without geometry	✓	✓	–	✓	46.01	52.47	44.38
Without reflection	✓	✓	✓	–	47.95	55.95	47.62

In the reported run, diagnostic reset outperforms neutral reset by 3.66 percentage points, suggesting a potential benefit from diagnostic context. It exceeds ordinary continuation by 17.07 points. Neutral reset also improves over continuation by 13.41 points, highlighting the value of restoring visual context in this cohort. With diagnostic text held fixed, restoring the parent view raises accuracy from 64.63% to 71.95%, a gain of 7.32 points. This pattern supports the design rationale for executable recovery: reflection can guide a change of observation state rather than remain a textual critique, although the added benefit of diagnostic content is run-dependent. These are descriptive point estimates on eligible reflection-triggering states, not evidence of statistical significance. Appendix G documents variation across runs.

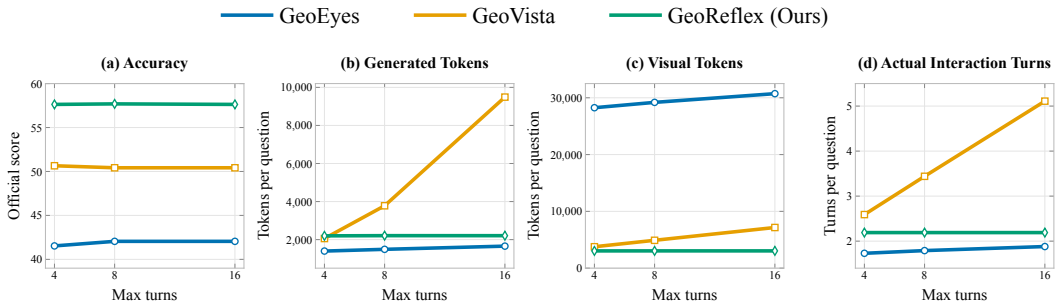


Figure 4: **Budget and cost on XLRS-Bench.** Panels show official accuracy and mean generated tokens, visual tokens, and interaction turns per question. GeoReflex maintains high accuracy and nearly constant resource use across budgets.

4.4 Interaction Budget and Inference Cost

We next compare the complete policies of GeoEyes, GeoVista, and GeoReflex-8B-Preview under interaction ceilings of four, eight, and sixteen turns (Fig. 4). Agents may stop earlier.

GeoReflex maintains approximately 57.6–58.0% accuracy with nearly constant resource use across budgets. At eight and sixteen turns, it outperforms GeoVista while using fewer generated tokens, visual tokens, and interaction turns; at four turns, it uses slightly more generated tokens. Compared with GeoEyes, GeoReflex achieves higher accuracy with fewer visual tokens but more generated tokens and interaction turns.

Table 3: Diagnosis and view restoration under a shared budget. Final-answer accuracy on 82 eligible XLRS questions from one run. Checkmarks denote enabled interventions.

Intervention	Add diagnosis	Restore view	Accuracy (%)
Without immediate restoration			
Continue	–	–	54.88
Text diagnosis	✓	–	64.63
With parent-view restoration			
Neutral reset	–	✓	68.29
Diagnostic reset	✓	✓	**71.95**

GeoReflex's advantage is therefore not contingent on a larger interaction ceiling. Strong accuracy with limited interaction is consistent with its recovery design: restoring the parent view provides a way to reconsider an unproductive crop instead of continuing along it. Together with the reward ablation, these results support recovery-oriented learning as a useful component of effective visual exploration. The budget sweep characterizes the complete policy rather than isolating the source of its resource savings; token counts remain proxies rather than FLOPs (Appendix G).

5 Conclusion

We introduced **GeoReflex**, a self-correcting active-perception framework, and **GeoReflex-48K**, a role-aligned corpus that pairs failed observation paths with diagnoses and corrective continuations. Using this corpus for supervised fine-tuning followed by reflection-efficient GRPO within our framework, we train **GeoReflex-8B-Preview**, a single 8B policy that performs visual exploration, reflection, and recovery without an external critic at inference. The resulting model achieves strong performance across three remote-sensing benchmarks: it improves over its SFT checkpoint on all three and surpasses the strongest compared external baselines on RSHR-Bench and XLRS-Bench by 5.40 and 7.04 points, respectively. It also maintains high accuracy with nearly constant resource use across interaction budgets. Reward ablation supports the contribution of reflection supervision, while matched-prefix controls highlight the value of parent-view restoration. Together, these results demonstrate the effectiveness of training visual agents to recover from incorrect observation paths using role-aligned data and recovery-oriented post-training.

Limitations. Reflection-trigger quality and the benefits of the data mixture remain independently unvalidated. Teacher diagnoses may be noisy, and data construction and reinforcement learning add computational overhead. Inference-cost evaluation is limited to XLRS-Bench and uses token counts rather than direct compute measurements.

AI USE STATEMENT

In this work, we used generative AI tools for drafting and editing manuscript text and for assisting with code related to data formatting checks. We have not used generative AI tools to fabricate experimental results. All scientific claims, protocol designs, and reported numbers are reviewed by the authors. We take responsibility for the final content of this work.

ETHICS STATEMENT

This work studies remote sensing vision–language agents on existing imagery and question–answer resources. We do not collect human subjects data. Potential dual-use concerns of high-resolution RS analysis are acknowledged; we focus on scientific VQA-style tasks and will follow dataset licenses upon release. Judge models used for offline rewrite may introduce labeling bias; we document taxonomy rules and plan human spot-checks.

REPRODUCIBILITY STATEMENT

We will release protocol specifications, GeoReflex-48K construction scripts, format validators, and evaluation configurations. Sections 3–3.3 describe the interaction protocol and learning objectives; Appendices C, E, and F provide construction and training details. The appendix includes complete benchmark results and evaluation protocol details.

REFERENCES

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. *arXiv preprint arXiv:2308.12966*, 2023.

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

Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, et al. Constitutional ai: Harmlessness from ai feedback. *arXiv preprint arXiv:2212.08073*, 2022.

Xinyun Chen, Maxwell Lin, Nathanael Schärli, and Denny Zhou. Teaching large language models to self-debug. In *International Conference on Learning Representations*, volume 2024, pp. 8746–8825, 2024.

Shehzaad Dhuliawala, Mojtaba Komeili, Jing Xu, Roberta Raileanu, Xian Li, Asli Celikyilmaz, and Jason Weston. Chain-of-verification reduces hallucination in large language models. In *Findings of the association for computational linguistics: ACL 2024*, pp. 3563–3578, 2024.

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*, 2025.

Luyu Gao, Zhuyun Dai, Panupong Pasupat, Anthony Chen, Arun Tejasvi Chaganty, Yicheng Fan, Vincent Zhao, Ni Lao, Hongrae Lee, Da-Cheng Juan, et al. Rarr: Researching and revising what language models say, using language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 16477–16508, 2023.

Zhibin Gou, Zhihong Shao, Yeyun Gong, Yujiu Yang, Nan Duan, Weizhu Chen, et al. Critic: Large language models can self-correct with tool-interactive critiquing. In *International Conference on Learning Representations*, volume 2024, pp. 57734–57811, 2024.

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

Jie Huang, Xinyun Chen, Swaroop Mishra, Huaixiu Steven Zheng, Adams Yu, Xinying Song, and Denny Zhou. Large language models cannot self-correct reasoning yet. In *International conference on learning representations*, volume 2024, pp. 32808–32824, 2024.

Wenxuan Huang, Bohan Jia, Shaosheng Cao, Zheyu Ye, Zhe Xu, Yao Hu, Shaohui Lin, et al. Vision-r1: Incentivizing reasoning capability in multimodal large language models. In *International Conference on Learning Representations*, volume 2026, pp. 63794–63812, 2026.

Ryo Kamoi, Yusen Zhang, Nan Zhang, Jiawei Han, and Rui Zhang. When can llms actually correct their own mistakes? a critical survey of self-correction of llms. *Transactions of the Association for Computational Linguistics*, 12:1417–1440, 2024.

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 *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 27831–27840. IEEE, 2024.

Aviral Kumar, Vincent Zhuang, Rishabh Agarwal, Yi Su, John D Co-Reyes, Avi Singh, Kate Baumli, Shariq Iqbal, Colton Bishop, Rebecca Roelofs, et al. Training language models to self-correct via reinforcement learning. In *International Conference on Learning Representations*, volume 2025, pp. 54523–54549, 2025.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023.

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.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, et al. Self-refine: Iterative refinement with self-feedback. *Advances in neural information processing systems*, 36:46534–46594, 2023.

Matthew Renze and Erhan Guven. Self-reflection in large language model agents: Effects on problem-solving performance. In *2024 2nd International Conference on Foundation and Large Language Models (FLLM)*, pp. 516–525. IEEE, 2024.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. Reflexion: Language agents with verbal reinforcement learning. *Advances in neural information processing systems*, 36:8634–8652, 2023.

Yan Shu, Bin Ren, Zhitong Xiong, Danda Pani Paudel, Luc Van Gool, Begüm Demir, Nicu Sebe, and Paolo Rota. Earthmind: Leveraging cross-sensor data for advanced earth observation interpretation with a unified multimodal llm. *arXiv preprint arXiv:2506.01667*, 2025.

Zhaochen Su, Peng Xia, Hangyu Guo, Zhenhua Liu, Yan Ma, Xiaoye Qu, Jiaqi Liu, Yanshu Li, Kaide Zeng, Zhengyuan Yang, et al. Thinking with images for multimodal reasoning: Foundations, methods, and future frontiers. *arXiv preprint arXiv:2506.23918*, 2025.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.

Di Wang, Shunyu Liu, Wentao Jiang, Fengxiang Wang, Yi Liu, Xiaolei Qin, Zhiming Luo, Chaoyang Zhou, Haonan Guo, Jing Zhang, et al. Geozero: Incentivizing reasoning from scratch on geospatial scenes. *arXiv preprint arXiv:2511.22645*, 2025.

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.

Fengxiang Wang, Jiangnan Huang, Mingshuo Chen, Yueying Li, Yang Shi, Junwei Luo, Haoyu Wang, Yansheng Li, Jing Zhang, Haiyan Zhao, et al. Beyond zooming: Learning multi-tool visual reasoning for ultra-high-resolution remote sensing. *arXiv preprint arXiv:2607.25993*, 2026c.

Jinwang Wang, Chang Xu, Wen Yang, and Lei Yu. A normalized gaussian wasserstein distance for tiny object detection. *arXiv preprint arXiv:2110.13389*, 2021.

Junjue Wang, Yanfei Zhong, Zihang Chen, Zhuo Zheng, Ailong Ma, and Liangpei Zhang. Earthvl: A progressive earth vision-language understanding and generation framework. *arXiv preprint arXiv:2601.02783*, 2026d.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024.

Sean Welleck, Ximing Lu, Peter West, Faeze Brahman, Tianxiao Shen, Daniel Khashabi, and Yejin Choi. Generating sequences by learning to self-correct. *arXiv preprint arXiv:2211.00053*, 2022.

Zhiheng Xi, Dingwen Yang, Jixuan Huang, Jiafu Tang, Guanyu Li, Yiwen Ding, Wei He, Boyang Hong, Shihan Do, Wenyu Zhan, et al. Enhancing llm reasoning via critique models with test-time and training-time supervision. *arXiv preprint arXiv:2411.16579*, 2024.

Zhihui Xie, Jie Chen, Liyu Chen, Weichao Mao, Jingjing Xu, and Lingpeng Kong. Teaching language models to critique via reinforcement learning. *arXiv preprint arXiv:2502.03492*, 2025.

Yong Xien Chng, Tao Hu, Wenwen Tong, Xueheng Li, Jiandong Chen, Haojia Yu, Jiefan Lu, Hewei Guo, Hanming Deng, Chengjun Xie, et al. Sensenova-mars: Empowering multimodal agentic reasoning and search via reinforcement learning. *arXiv e-prints*, pp. arXiv–2512, 2025.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. React: Synergizing reasoning and acting in language models. *arXiv preprint arXiv:2210.03629*, 2022.

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

Zilun Zhang, Haozhan Shen, Tiancheng Zhao, Zian Guan, Bin Chen, Yuhao Wang, Xu Jia, Yuxiang Cai, Yongheng Shang, and Jianwei Yin. Imagerag: Enhancing ultra high resolution remote sensing imagery analysis with imagerag. *arXiv preprint arXiv:2411.07688*, 2024.

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.

Ziwei Zheng, Minghao Yang, Jack Hong, Chenxiao Zhao, Guohai Xu, Le Yang, and Chao Shen. Deepeyes: Incentivizing" thinking with images" via reinforcement learning. In *International Conference on Learning Representations*, volume 2026, pp. 126775–126798, 2026.

Yunqi Zhou, Chengjie Jiang, Chun Yuan, and Jing Li. Look where it matters: Training-free ultra-hr remote sensing vqa via adaptive zoom search. *arXiv preprint arXiv:2511.20460*, 2025.

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.

A Additional protocol details

Interaction implementation. For task x and state h_i defined in Sec. 3, each model call is implemented as follows:

$$p_i = \mathcal{E}(x, h_i), \qquad y_i \sim \pi_\theta(\cdot \mid p_i), \qquad a_i = \mathcal{P}(y_i), \qquad (h_{i+1}, o_i) = \mathcal{T}(h_i, y_i, a_i). \tag{12}$$

$\mathcal{E}$ constructs the multimodal input p_i from the task and current state. The policy generates response y_i, which $\mathcal{P}$ interprets as an action a_i or a final answer. The transition $\mathcal{T}$ executes that decision, returns observation o_i (empty for a terminal answer), and updates the state. Thus later reasoning is conditioned on the evidence actually acquired, rather than on the initial image alone. The parser and transition preserve one action per response. Adjacent call inputs need not be concatenated into a single token sequence. The intended continuation u_i changes the recovery observation wording but not the action space; a nominal ANSWER may still inspect another view.

The default GeoReflex evaluation protocol enforces one action per turn, a maximum of eight turns and three recovery actions. Training and evaluation share state-transition semantics, but their turn limits need not coincide; training limits are specified below. The budget sweep in Sec. 4.4 separately varies the interaction ceiling to four, eight, and sixteen turns. If a benchmark box dimension exceeds 1000 pixels, the wrapper normalizes coordinates using the stated image resolution when available. Crops shorter than 224 pixels on either side are bilinearly resized to approximately 224 pixels without enlarging the box or adding context. A matched single-turn control disables visual actions.

Recovery boundary conditions. In the evaluation environment, an accepted fallback pops the current view only when the stack contains a parent. At the root, it leaves the view unchanged but still emits the recovery observation and consumes one recovery action. Once the recovery budget is exhausted, the environment returns a rejection observation without changing the view or incrementing the recovery count. The rejection asks the policy to answer on the current view; it is not itself a forced terminal answer.

Baseline provenance and comparability. All scores in the main and complete benchmark comparisons are obtained with our own evaluation runs except for GeoLens on RSHR and XLRS, whose scores are quoted from Wang et al. (2026c). GeoLens on MME-RS is evaluated by us. For our main runs, interactive methods have an eight-turn ceiling; tools, model-specific preprocessing, and other evaluation parameters follow their official implementations. Benchmark answer scoring and aggregation use the official evaluation programs. The quoted GeoLens results retain their source protocol and are not claimed to have been reproduced under our eight-turn ceiling. A shared interaction ceiling does not imply identical visual-token budgets or computation across methods.

Resolution conventions. Source-image resolution, global-view resolution, individual tile size, and the total visual-input budget are distinct quantities. We retain each baseline's official preprocessing rather than resize all methods to a common square input. Dynamic resolution denotes variable image dimensions subject to the backbone's preprocessing and resource limits, not unlimited input resolution. In particular, a fixed tile size does not specify the total number of visual tokens, and a resolution interval used in training does not by itself establish a hard inference limit.

B REFLECTION JUDGE PROTOCOL

The training judge uses Qwen3-VL-32B-Thinking. The following summarizes the system and user prompts implemented in our training code; it is not a verbatim prompt transcription. The judge evaluates the attempt without rewriting the student and returns one JSON object. Its reference-answer access is confined to training-time assessment.

Error taxonomy and necessity. The prompt distinguishes seven primary error categories. Localization, counting, and hallucination have high recoverability; spatial relation, context integration, and attribute errors have medium recoverability; fine-grained recognition has low recoverability. Localization concerns an incorrect location after establishing the target, whereas an unsupported premature crop is categorized as hallucination. Counting includes missed or duplicated objects and a plan without a numerical answer. Context integration concerns inconsistent aggregation of otherwise plausible local evidence. One primary code is preferred, with at most two when both are independently necessary. The judge assigns reflection necessity only to high- or medium-recoverability errors that reflection could still address. An already-correct path or fine-grained recognition alone does not warrant reflection.

Observed action and value scores. The parser supplies `did_reflect`; the judge must not override it. A crop alone is not reflection. The diagnosis score evaluates whether the reflection names the actual mistake using process-level language; inserting reference-answer values into student reflection is treated as a defect. The action score evaluates the fallback reason, requested continuation, and box against the restored context. The improvement score measures whether reflection at this step is expected to move the answer toward the reference, using the subsequent recovery turn when available. Diagnosis and action receive zero when no reflection occurred; improvement receives zero when reflection is not warranted. All raw scores lie in $[0, 1]$ and are mapped as in Eq. 11.

The prompt treats `after=answer` as a request to commit on the restored view and `after=replan` as permission to inspect further. These are judge criteria, not hard environment constraints: the environment still permits later visual actions in either case. For multiple fallback events, the input builder preferentially uses the last fallback's reflection text and arguments while also supplying the action trace and trajectory. The resulting reward is a trajectory-level assessment, not an average of independently judged recovery events.

User-prompt contents and output schema. The user prompt supplies task type, question, reference and predicted answers, the detected reflection flag, crop and fallback counts, the last action, student reflection text, and fallback arguments. These fields are followed by the parsed action summary, the reconstructed reflection-round input, the subsequent recovery turn when available, and the student trajectory. Missing recovery is explicitly represented as absent rather than as a successful correction. The required output fields are:

```
{``should_reflect'': boolean, ``error_types'': [codes],
``diagnosis'': number, ``action'': number,
``improvement'': number, ``confidence'': number,
``rationale'': string}
```

Confidence is bounded to $[0, 1]$. The rationale contains one to three sentences and may mention the reference answer because it is a reward log, not a student rewrite. Confidence, rationale, and error codes are not additional terms in Eq. 11.

C DATA CONSTRUCTION CHECKLIST

Format checks enforce one action per assistant turn, align every visual observation with its trajectory state, and preserve the executed failure prefix in both reflection and refinement samples. The offline rewrite is an intermediate only; training uses the recovery transition formalized in Eq. 4 (Sec. 3.2). The final corpus contains $47{,}655$ SFT examples. Qualitative reflection and refinement examples are taken from truncated ShareGPT samples in the data kit.

Sampling and teacher annotation. The Generator draws from approximately $4{,}446$ questions in the APE-GRO splits, sampling at most eight Qwen3-VL-8B-Thinking rollouts per question and eight interaction turns per rollout. During rejection sampling for initial data construction, visual-grounding attempts are classified as successful when

$$\tanh\left(\frac{100}{d_{\text{center}} + 80}\right) > 0.4, \tag{13}$$

where d_{center} is the predicted-to-target box-center distance. This criterion is used only for data selection, not for benchmark evaluation or reinforcement-learning rewards. Benchmark results use the official evaluation functions described in Sec. 4.1. The data-construction pool includes 564 mixed-outcome questions and 1,770 questions without a successful student attempt. In one call, the teacher diagnoses a failed attempt using seven remote-sensing error categories and writes exactly one recovery turn. Approximately $16{,}418$ source candidates yield $13{,}771$ candidate reflection–refinement pairs before final filtering.

Final mixture. After filtering and deduplication, $4{,}492$ pairs cover $1{,}034$ questions. We impose no maximum number of pairs per question. The final corpus contains $33{,}445$ teacher examples (70.2%), $5{,}226$ student successes (11.0%), and $4{,}492$ examples for each recovery role (9.4% each). Student successes comprise $3{,}976$ active-interaction trajectories and $1{,}250$ direct answers.

D DETAILED BENCHMARK RESULTS

Table 4: XLRS-Bench results by perception and reasoning category. Best and second-best results are bold and underlined; Avg. is the official benchmark score.

Method	Perception				Reasoning				Avg.
	Counting	Scene Class.	Object Spatial Relationship	Object Properties	Complex Reasoning	Planning	Spatiotemporal Reasoning	Anomaly Reasoning	
Closed-source MLLMs									
GPT-4o	29.51	48.55	32.35	24.78	52.60	41.24	21.85	72.06	32.15
Gemini-3.8-flash	**45.00**	47.00	38.80	38.67	63.00	35.00	38.33	72.00	42.37
Claude-4.5-Sonnet	31.25	51.00	28.60	35.30	52.00	42.00	11.67	74.00	37.63
Open-source MLLMs									
DeepSeek-VL2-small	41.25	51.67	31.60	36.69	60.00	37.00	46.67	69.00	40.32
LLaVA1.5-7B	30.00	39.67	25.40	34.76	49.50	35.00	20.00	64.00	35.10
LLaVA-UHD-v3	27.50	49.33	20.80	36.93	48.50	29.00	26.67	74.00	36.53
LLaVA-OneVision-2	33.75	48.00	28.60	33.92	69.50	38.00	44.67	**78.00**	38.54
Large-Format MLLMs									
Qwen3-VL-8B-Instruct	36.88	51.00	31.20	42.11	64.00	32.00	45.00	74.00	43.12
Qwen3-VL-8B-Thinking	33.12	60.67	26.00	39.76	61.00	39.00	38.33	72.00	41.59
GLM-4.6V-Flash	<u>43.13</u>	53.33	38.00	38.19	66.50	48.00	46.67	74.00	43.38
InternVL3.5-8B	35.63	52.67	<u>39.00</u>	35.66	65.50	50.00	48.33	73.00	41.72
Qwen2.5-VL-32B-Instruct	38.12	59.00	31.40	37.23	64.00	52.00	35.00	72.00	41.75
Remote Sensing MLLMs									
GeoLLaVA-8K	30.00	54.00	24.20	32.17	62.00	<u>61.00</u>	46.67	65.00	37.11
GeoChat	23.75	31.33	23.80	22.17	39.00	49.00	35.00	56.00	26.72
VHM	19.38	25.33	29.40	29.22	35.50	23.00	35.00	55.00	29.51
EarthDial	33.12	21.33	26.20	32.35	60.50	27.00	<u>55.00</u>	61.00	33.34
Methods									
ZoomEarth	33.12	37.67	22.60	33.73	55.50	38.00	30.00	66.00	34.81
ZoomSearch	36.25	48.33	28.80	39.52	**81.00**	33.00	45.00	<u>76.00</u>	40.68
ImageRag	39.37	55.67	26.20	36.93	68.00	33.00	41.67	70.00	40.19
Coarse-to-fine	35.00	35.33	26.80	33.37	<u>76.00</u>	24.00	**58.33**	69.00	35.65
GeoEyes	32.50	51.67	31.80	40.72	61.50	43.00	41.67	71.00	42.34
DeepEyes	30.63	48.33	26.40	33.86	49.50	39.00	36.67	60.00	35.97
RS-EoT	41.25	46.67	23.80	36.02	58.00	46.00	1.66	58.00	37.14
RSCoVLM	29.38	38.67	36.60	43.55	63.00	45.00	15.00	74.00	42.95
GeoLens	38.13	<u>63.33</u>	36.80	46.73	68.50	**62.00**	51.70	72.00	49.12
GeoVista-7B-Preview	41.25	54.00	**42.40**	52.47	52.50	46.00	48.33	69.00	<u>50.65</u>
Ours									
GeoReflex-8B-Instruct	31.25	40.00	26.80	<u>59.46</u>	54.50	43.00	48.33	65.00	49.90
GeoReflex-8B-Preview	39.37	**65.67**	35.80	**64.16**	64.00	48.00	43.33	71.00	**57.69**

Table 5: RSHR-Bench task-level results. Reasoning: AD=Anomaly, FP=Future Prediction, MRJC and MRJCS=Multi-region Joint Contrast, OSJ=Object State Judgment. Perception: COL=Color, DET=Detection, SHP=Shape, OC=Classification, REL=Spatial Relationship, OCN and RCN=Counting, OGD and RG=Grounding. Multi-turn: MTFP=Future Prediction, MAD=Anomaly, MOSJ=Object State Judgment.

Model	Reasoning					Reas. Avg.	Perception									Perc. Avg.	Multi-turn			MTEM @1
	AD	FP	MRJC	MRJCS	OSJ		COL	DET	SHP	OC	REL	OCN	RCN	OGD	RG		MTFP	MAD	MOSJ	
Closed-source MLLMs																				
GPT-4o	68.0	**56.0**	30.5	32.0	64.0	50.1	49.5	15.0	23.0	35.5	30.5	22.5	41.0	28.0	27.1	30.2	72.0	70.0	84.1	47.4
Gemini-3.8-flash	70.0	38.0	50.0	26.0	68.0	50.5	55.5	24.0	56.0	25.0	44.5	20.0	40.0	34.0	44.3	37.2	**74.0**	70.0	83.3	**52.6**
Claude-4.5-Sonnet	60.0	38.0	15.0	40.0	64.0	47.3	40.5	19.0	24.0	19.0	41.5	25.5	34.0	31.0	35.7	30.4	67.3	70.0	75.0	40.4
Open-source MLLMs																				
DeepSeek-VL2-small	36.0	28.0	25.0	28.0	66.0	38.2	52.0	29.0	32.0	29.0	50.5	23.0	33.0	27.0	27.1	34.7	63.3	63.3	69.7	26.3
LLaVA1.5-7B	40.0	28.0	30.0	30.0	46.0	35.5	51.0	24.0	20.0	29.5	27.0	18.0	24.0	25.5	22.9	28.2	46.0	60.0	48.5	14.0
LLaVA-UHD-v3	38.0	28.0	20.0	44.0	44.0	36.8	45.0	31.0	27.0	51.5	46.0	32.0	12.0	24.5	30.0	35.7	64.7	65.0	62.1	27.2
LLaVA-OneVision-2	38.0	32.0	15.0	36.0	58.0	38.6	41.0	35.0	24.0	28.0	52.5	23.5	34.0	34.5	37.1	34.9	64.7	70.0	74.2	31.6
Large-Format MLLMs																				
Qwen3-VL-8B-Instruct	72.0	46.0	50.0	32.0	66.0	53.6	56.0	29.0	23.0	28.5	47.0	15.0	47.0	37.0	**47.1**	36.4	70.0	**80.0**	82.6	46.5
Qwen3-VL-8B-Thinking	**74.0**	42.0	10.0	28.0	62.0	47.7	54.0	16.0	22.0	32.5	51.0	17.0	46.0	36.5	22.9	35.2	70.0	70.0	78.0	43.0
GLM-4.6V-Flash	**74.0**	38.0	45.0	40.0	62.0	52.7	51.5	28.0	16.0	38.5	52.5	21.5	47.0	**40.0**	44.3	38.7	64.0	71.7	64.4	35.1
InternVL3.5-8B	70.0	50.0	35.0	40.0	60.0	53.2	46.5	**39.0**	25.0	46.5	19.5	24.5	33.0	32.5	**47.1**	34.2	66.7	78.3	82.6	45.6
Qwen2.5-VL-32B-Instruct	58.0	**56.0**	**55.0**	42.0	50.0	51.8	43.5	28.0	17.0	35.0	50.0	22.0	38.0	23.5	31.4	33.1	64.7	73.3	59.9	31.6
Remote Sensing MLLMs																				
GeoLLaVA-8K	46.0	50.0	40.0	40.0	54.0	46.8	45.5	28.0	27.0	22.5	10.5	**38.5**	34.0	27.0	27.1	28.9	60.7	50.0	74.2	26.3
GeoChat	42.0	26.0	45.0	24.0	46.0	35.6	39.5	24.0	22.0	34.0	27.5	21.5	34.0	24.0	22.9	28.4	50.7	58.3	56.8	13.2
VHM	26.0	28.0	15.0	26.0	32.0	26.8	36.0	26.0	21.0	37.0	31.0	8.5	20.0	27.0	37.1	27.2	49.3	51.7	43.2	12.3
EarthDial	56.0	38.0	40.0	38.0	56.0	46.4	37.5	22.0	28.0	28.5	27.0	17.0	34.0	32.5	28.6	28.4	60.7	55.7	68.9	29.0
Methods																				
ZoomEarth	50.0	42.0	40.0	38.0	46.0	43.6	44.0	29.0	17.0	21.5	40.0	23.0	31.0	26.0	20.0	29.2	56.0	66.7	56.1	22.8
ZoomSearch	54.0	30.0	45.0	36.0	66.0	46.4	42.0	28.0	26.0	20.5	41.5	17.5	36.0	25.5	21.4	29.1	58.7	63.3	72.0	28.1
ImageRag	72.0	**56.0**	**55.0**	**48.0**	66.0	**60.0**	51.0	22.0	18.0	20.5	44.0	14.5	34.0	35.0	22.9	30.7	62.7	73.3	80.3	34.2
Coarse-to-fine	26.0	22.0	35.0	26.0	30.0	26.8	27.0	28.0	33.0	27.0	25.5	25.0	29.0	23.5	30.0	26.8	21.3	28.3	25.8	2.6
GeoEyes	64.0	42.0	45.0	34.0	60.0	49.6	56.5	34.0	25.0	36.0	55.0	19.5	34.0	36.5	32.9	38.2	64.7	60.0	81.1	38.6
DeepEyes	50.0	44.0	40.0	28.0	58.0	44.6	74.5	26.0	34.0	58.0	52.5	25.0	32.0	31.0	25.7	43.2	63.3	53.3	72.0	28.1
RS-EoT	48.0	26.0	40.0	32.0	62.0	41.8	49.0	19.0	20.0	31.5	38.5	18.0	28.0	23.5	21.4	29.4	37.3	35.0	40.9	10.5
RSCoVLM	62.0	40.0	40.0	34.0	**72.0**	50.9	48.0	18.0	21.0	31.0	34.0	21.5	23.0	27.5	**47.1**	30.6	61.3	65.0	75.8	32.5
GeoLens	68.0	46.0	50.0	42.0	62.0	53.6	58.5	27.0	22.7	35.0	**57.5**	34.5	39.0	33.0	38.6	38.4	61.1	70.0	**85.4**	48.8
GeoVista-7B-Preview	54.0	18.0	40.0	32.0	52.0	39.1	68.0	34.0	50.0	60.5	28.0	26.5	31.5	34.0	32.9	41.8	62.7	68.3	78.8	36.8
Ours																				
GeoReflex-8B-Instruct	50.0	28.0	35.0	30.0	50.0	39.1	72.0	26.0	**60.0**	56.0	40.5	20.0	35.0	26.5	27.1	41.6	51.3	58.3	58.3	14.9
GeoReflex-8B-Preview	64.0	30.0	35.0	36.0	56.0	45.5	**82.0**	36.0	**60.0**	**67.5**	51.0	21.5	**50.0**	29.0	40.0	**49.3**	60.7	65.0	74.2	26.3

E TRAINING CONFIGURATION

Supervised fine-tuning. We fine-tune Qwen3-VL-8B-Thinking on 47,655 examples with LoRA on all target modules ($r = 16$, $\alpha = 32$, dropout 0.05). Training uses two GPUs, a 16,384-token cutoff, maximum image area 1,016,064 pixels, bf16, FlashAttention 2, and gradient checkpointing. We train for two epochs with learning rate 5×10^{-5}, cosine decay, warmup ratio 0.1, weight decay 0.01, gradient norm 1.0, and effective batch size 32. The SFT run takes 2,980 steps and 11 h 18 min 24 s, reaching training loss 0.2289.

Supervised loss masking. SFT uses LlamaFactory with multiturn ShareGPT examples. Prompt training is disabled: system and user messages, tool observations represented as user messages, and padding are masked from the loss. All assistant turns are supervised, rather than only the final response. In the paired recovery examples, retained assistant messages in the failure prefix also contribute to the loss. Role alignment specifies the newly constructed target turns, not exclusive role-specific loss masks.

GRPO configuration. GRPO starts from the SFT checkpoint at step 2,980, uses four prompts per batch and four rollouts per prompt, and applies a learning rate of 10^{-6}. The actor minibatch contains 16 rollouts. KL penalties in both reward and loss are disabled, as are entropy regularization and length penalties. The training configuration caps assistant and user turns at five each; this is distinct from the default eight-call evaluation ceiling. The policy loss uses token-mean aggregation and dual clipping, as specified in Appendix F.

Reward configuration. The full-reward configuration uses $(\omega_{\text{fmt}}, \omega_{\text{acc}}, \omega_{\text{geo}}, \omega_{\text{ref}}) = (0.1, 1.0, 0.8, 0.7)$, $\gamma_{\text{fmt}} = 0.8$, $C_B = 20$, counting distance scale 2, and $\lambda_{\text{nwd}} = 0.5$. Consequently, for one grounding box the combined accuracy and geometry contribution is $1.4\,\text{NWD} + 0.4\,\text{IoU}$. The no-geometry variant retains the accuracy NWD term. Reflection agreement is $+1$ when the judge's necessity label matches detected reflection and -1 otherwise; raw component scores are clipped to $[0, 1]$ before affine mapping. The GRPO entry points explicitly disable the shared reward module's length-penalty option.

Counting and multi-box rewards. Box NWD is defined in Eq. 9; the following specifies counting and multi-box aggregation.

For scalar counts $\hat{n}$ and n, define $\sigma(n) = \max(|n|, 1)^{1/3}$ and

$$r_{\text{count}} = \exp\left(-\frac{\sqrt{(\hat{n}-n)^2 + (\sigma(\hat{n}) - \sigma(n))^2}}{C_n}\right), \qquad C_n = 2. \tag{14}$$

For box sets, candidate pairs are ranked by their similarity and greedily matched one-to-one. Matched scores are summed and divided by the larger set cardinality, so unmatched boxes reduce the reward. Accuracy matching uses NWD; geometry matching independently uses the NWD–IoU mixture in Eq. 10. Either empty set yields zero box reward. When scalar and box-based counting scores are both available, accuracy is their equally weighted mean; otherwise it uses the available score, or zero if neither can be computed. Categorical answers are compared after trimming whitespace and converting to uppercase.

F GRPO OPTIMIZATION DETAILS

Group-relative, token-level optimization. We convert rollout rewards into group-normalized advantages

$$\hat{A}_k = \frac{R(\tau_k) - \text{mean}_j\, R(\tau_j)}{\text{std}_j\, R(\tau_j) + \epsilon_{\text{std}}}, \tag{15}$$

where $\epsilon_{\text{std}} > 0$ ensures numerical stability. To preserve the multi-call conditioning in Eq. 12, let $\mathcal{I}_k = \{(i,t) : t \text{ indexes a policy-generated token in call } i\}$ and $N_k = |\mathcal{I}_k|$. For call input $p_{k,i}$ and generated token $y_{k,i,t}$, the policy ratio is

$$\rho_{k,i,t}(\theta) = \frac{\pi_\theta(y_{k,i,t} \mid p_{k,i}, y_{k,i,<t})}{\pi_{\theta_{\text{old}}}(y_{k,i,t} \mid p_{k,i}, y_{k,i,<t})}. \tag{16}$$

The rollout advantage $\hat{A}_k$ is shared across its generated tokens. Let $\mathcal{B}$ index the rollouts in an actor update minibatch, with advantages computed within their respective prompt groups. Define the clipped surrogate

$$s_{k,i,t} = \min\left[\rho_{k,i,t}\hat{A}_k,\ \text{clip}(\rho_{k,i,t}, 1-\varepsilon, 1+\varepsilon)\hat{A}_k\right]. \tag{17}$$

The dual-clipped surrogate and token-mean loss are

$$\tilde{s}_{k,i,t} = \begin{cases} \max(s_{k,i,t}, c_{\text{dual}}\hat{A}_k), & \hat{A}_k < 0, \\ s_{k,i,t}, & \hat{A}_k \geq 0, \end{cases} \qquad \mathcal{L}_{\text{GRPO}} = -\frac{\sum_{k\in\mathcal{B}}\sum_{(i,t)\in\mathcal{I}_k}\tilde{s}_{k,i,t}}{\sum_{k\in\mathcal{B}} N_k}. \tag{18}$$

The clipping range is $\varepsilon = 0.2$ and the implementation's default dual-clip constant is $c_{\text{dual}} = 3$. For numerical stability, log probability ratios are clamped to $[-20, 20]$ before exponentiation. Task inputs and tool observations are excluded through the response mask. Token-mean aggregation weights generated tokens equally, rather than giving equal weight to trajectories of different lengths. No KL term or learned value model is used.

G ADDITIONAL EVALUATION DETAILS

Reward-ablation execution. Table 2 reports completed results for both reduced-reward variants with unchanged tool availability. Removing geometry tests the additional geometric reward while retaining the grounding NWD accuracy signal; removing reflection tests reflection supervision beyond task outcomes. Scores use official evaluation functions. These single-seed comparisons do not establish stability across training runs.

Inference-only mechanism controls. Table 3 reports the completed comparison on 82 XLRS-Bench trajectories in which GeoReflex triggers reflection. The four continuations start from matched prefixes preceding the reflection decision. Continue retains the current view without an injected diagnosis; Text diagnosis supplies a diagnosis without an immediate reset; Neutral reset restores the parent view with a neutral note; Diagnostic reset restores the parent view with a student diagnosis. The analysis addresses the contribution of diagnosis to recovery from matched prefixes.

Cohort and matched prefixes. The cohort contains all 82 eligible questions obtained from the 3,080-question XLRS evaluation, covering 80 source images. Selection uses the first assistant turn containing a reflection marker or fallback reference, and freezes the history before that turn. Eligibility requires a current crop view, at least one preceding zoom, an unexhausted fallback budget, at least two remaining model calls under the eight-call ceiling, and a resolvable image path. Selection does not depend on final correctness. The cohort covers 2.66% of benchmark questions; this is eligibility coverage, not overall reflection incidence.

A diagnosis is generated once from each frozen prefix and cached for both diagnosis conditions. It is not the original trajectory's reflection text, which is stored separately. Diagnosis generation consumes one model call and its tokens are charged to the corresponding condition. Parent restoration counts as a tool call. All continuations retain access to zoom, fallback, and answering. All four scores in Table 3 belong to the same run. A separate archived run under the same settings yields 64.63% for diagnostic reset rather than 71.95%, while the other three conditions have the same aggregate scores. The 64.63% result is lower than neutral reset (68.29%), so the incremental benefit of diagnosis is not consistent across these two runs. No paired uncertainty estimates are reported; these comparisons are descriptive.

Scope of the recovery cohort. Because the cohort is conditioned on the policy having triggered reflection, it does not measure missed reflection opportunities or unnecessary triggers across the full benchmark. Autonomous trigger quality requires an independently annotated set of decision states.

Inference-cost interpretation. Section 4.4 compares GeoEyes, GeoVista, and GeoReflex on XLRS-Bench at four, eight, and sixteen turns. It reports official scores, generated tokens, visual tokens, and logged interaction turns; the latter are not tool-call counts. The comparison characterizes observed performance and resource use, not a controlled ablation of reflection reward. Cross-model token counts are not directly equivalent to computation, and latency is not measured.

Evidence limitations and qualitative examples. Teacher diagnoses and training-judge scores have not yet been independently validated by a human audit. The current results therefore do not establish the reliability of every diagnosis or the absence of hindsight bias. Appendix H presents selected evaluation trajectories, including imperfect intermediate descriptions, rather than a quantitative audit of supervision quality.

H REPRESENTATIVE EVALUATION CASES

Figures 5–8 present four representative cases selected during evaluation. Each panel shows the question, visual context, and the model's inspection, reflection, fallback, and final response. The selected trajectories illustrate different forms of reconsideration; they are qualitative examples rather than estimates of recovery frequency or verification of every intermediate claim.

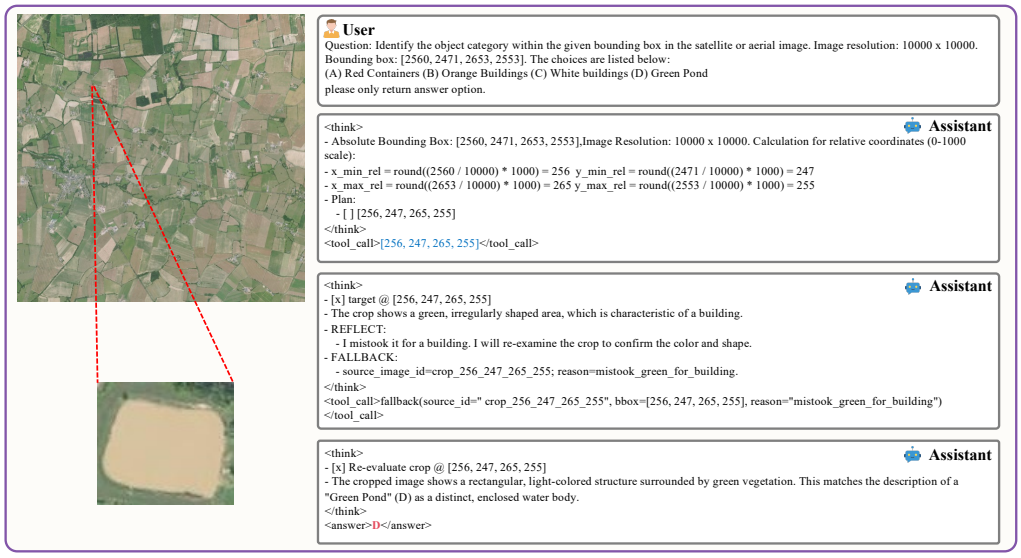


Figure 5: **Object properties.** Fallback revises the initial building interpretation to the pond option, although the continuation retains imprecise descriptions. A revised answer need not imply fully consistent reasoning.

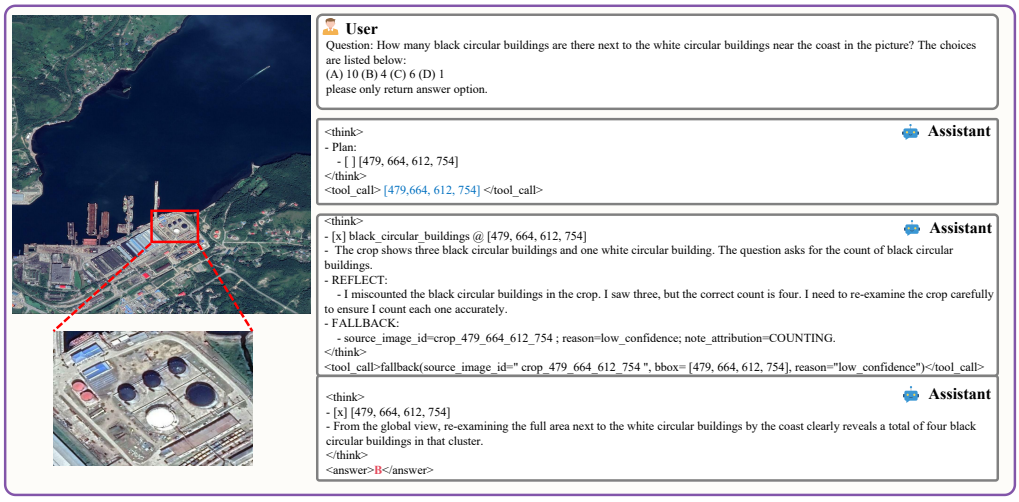


Figure 6: **Counting.** After diagnosing a counting error and invoking fallback, the model revises three black circular buildings to four and selects B.

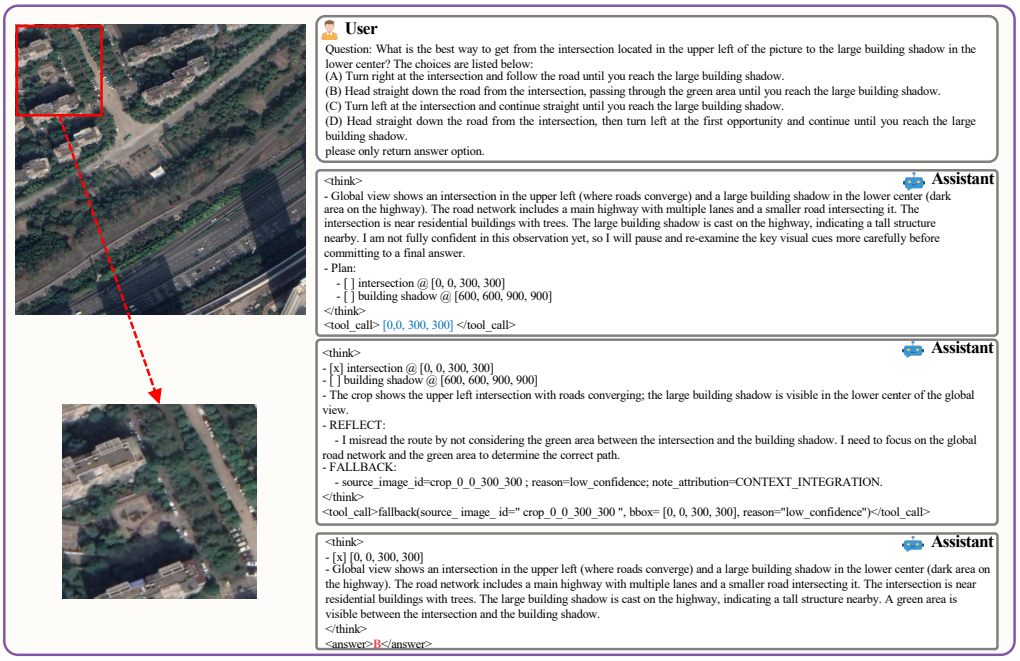


Figure 7: **Route planning.** Recognizing missing green-area context prompts fallback from the intersection to the global road layout, followed by answer B.

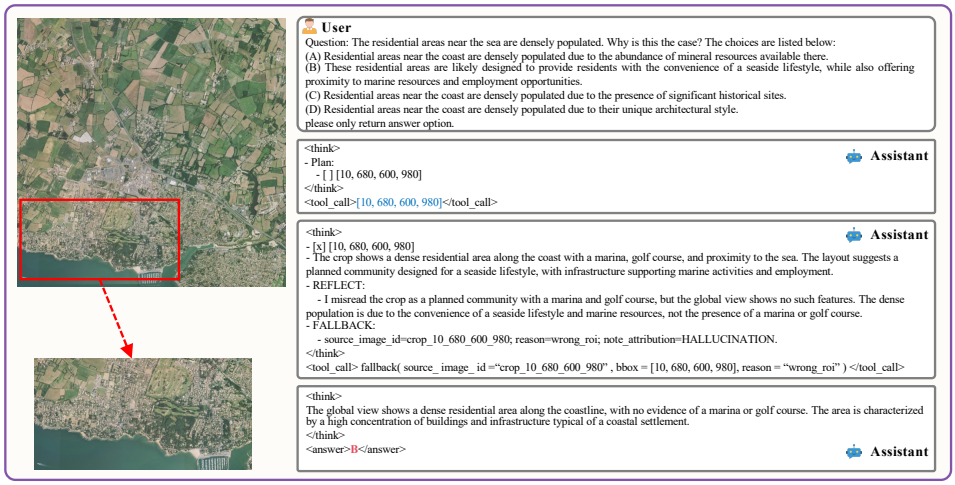


Figure 8: **Complex reasoning.** The model retracts claims about a marina and golf course, invokes fallback, and selects B. This shows evidence revision, not a verified causal explanation of population density.