REDIS: REASONING–DIAGNOSIS SELF-EVOLUTION FOR MULTIMODAL REASONING IN REMOTE SENSING

Anonymous authors
Paper under double-blind review

ABSTRACT

Multimodal large language models (MLLMs) have shown strong potential for remote sensing understanding, yet reliable reasoning remains challenging when tasks require fine-grained visual grounding and multi-step inference. Existing approaches typically rely on either costly pre-constructed reasoning trajectories or self-generated signals such as response consistency, which can reinforce shared errors and provide unreliable supervision. We introduce **Reasoning–Diagnosis Self-Evolution (ReDiS)**, a framework that separates reasoning exploration from candidate evaluation while allowing the two processes to evolve jointly. ReDiS instantiates two role-specific experts from a shared multimodal backbone: a **Reasoner** that explores diverse visual evidence–answer candidates and a **Diagnoser** that jointly evaluates the complete candidate set and provides credibility-aware feedback. As the Reasoner evolves, it exposes new candidate distributions and failure modes; the Diagnoser adapts to these changes and, in turn, reshapes the learning signal for subsequent exploration. To handle groups in which all sampled answers are incorrect and group-relative advantages vanish, we further introduce **Failure-Triggered Recovery (FTR)**, which provides minimal outcome-level correction without prescribing intermediate reasoning trajectories. Experiments on LHRS-Bench, MME-RS, and CHOICE show consistent improvements over Qwen3.5 backbones at 0.8B, 2B, and 4B scales. ReDiS improves category-averaged accuracy by 6.47 percentage points at 0.8B, while ReDiS-4B achieves the highest average performance among the compared remote sensing MLLMs. Code and data will be released publicly.

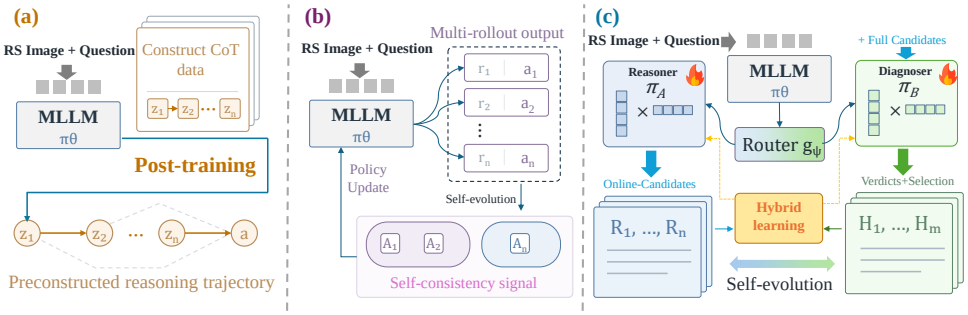

Figure 1: Training paradigms for remote sensing reasoning. (a) Trajectory-supervised methods learn from pre-constructed chains of thought. (b) Self-consistency methods derive feedback from multiple online rollouts. (c) The Reasoning–Diagnosis Self-Evolution (ReDiS) framework for multimodal reasoning instantiates two role-specific experts from a shared multimodal model.

1 INTRODUCTION

MLLMs (Yang et al., 2025; Dai et al., 2023; Chen & Xing, 2024; Liu et al., 2023) have advanced remote sensing interpretation, from image captioning (Cheng et al., 2022; Wang et al., 2026a) and visual question answering (Lobry et al., 2020; Wang et al., 2024; Chappuis et al., 2022) to

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.

visual grounding and geospatial reasoning (Yao et al., 2026a; Liu et al., 2026). However, reliable reasoning remains challenging in scenes with small, densely distributed objects and complex spatial relationships (Yu et al., 2006). Answering questions in these scenes requires identifying subtle visual evidence and integrating multiple observations. Errors in recognizing objects, attributes, or relations can propagate through subsequent inference, producing plausible answers based on unsupported visual premises. Reliable remote sensing interpretation therefore requires both correct answers and reasoning grounded in question-relevant visual evidence.

This exposes the limitation of simple answer-level supervision: it can verify whether a final prediction is correct, but provides little information about the reasoning process. A common approach is to construct reasoning trajectories and use them as intermediate supervision (Yao et al., 2026a). Such trajectories provide richer guidance, but high-quality traces are expensive to generate and filter, and they are fixed before policy optimization. As the policy evolves, its errors and hallucination patterns may no longer be well covered by these pre-constructed traces. Self-evolution methods obtain feedback from online rollouts of the current policy (Thawakar et al., 2025). Although these rollouts reflect evolving model behavior, agreement among them need not indicate visual correctness: multiple responses may reinforce the same visual misconception (Venkatraman et al., 2026). Recent methods strengthen visual feedback through invariance constraints (Venkatraman et al., 2026) or region-level verification (Yao et al., 2026b). These advances motivate a complementary question: ***how can reasoning exploration and joint candidate diagnosis evolve together while keeping self-generated feedback anchored to task correctness?***

To address this, we take a different design, which separates exploration from evaluation while allowing them to evolve together. One process explores alternative visual evidence and answers, while the other compares the resulting candidates and judges their reliability. As exploration evolves, it continually changes the candidate distribution and failure modes that evaluation must learn to distinguish. The evaluator, in turn, changes the learning signal seen by exploration, reshaping which behaviors are reinforced in the next round. Their evolution remains constrained by answer-level supervision, which fixes the criterion of correctness outside this loop.

We introduce ReDiS, a Reasoning–Diagnosis Self-evolution framework built around two role-specific experts: a Reasoner and a Diagnoser. Figure 1 contrasts ReDiS with trajectory-supervised and self-consistency paradigms. Given an image–question pair, the Reasoner samples a set of evidence–answer candidates from its current policy. The Diagnoser observes the image, question, and the complete candidate set, assigns a diagnostic state to each candidate, and either selects a preferred response or rejects the entire set. After the rollouts are generated, the reference answer is used to construct correctness labels for the Reasoner and answer-anchored targets for the Diagnoser; both roles are then optimized with group-relative objectives. The Diagnoser's judgments are further aggregated into candidate-level credibility scores, which refine the Reasoner's reward only for answer-correct candidates, while incorrect candidates receive no positive reward regardless of the diagnosis. A remaining failure case arises when exploration produces no correct candidate. In this case, correctness-gated rewards collapse across the group, leaving group-relative optimization with no direction to improve. We address this with Failure-Triggered Recovery (FTR), which intervenes only on failed groups and supplies minimal supervision to the answer value tokens. FTR therefore complements the reasoning–diagnosis loop: diagnosis guides learning when exploration is informative, while minimal correction is used only when exploration provides no usable signal.

Experiments on three remote sensing visual question answering (RSVQA) benchmarks, including LHRS-Bench, MME-RS, and CHOICE, show consistent improvements over the Qwen3.5-0.8B, 2B, and 4B backbones. At 0.8B, ReDiS improves average performance by 6.47 percentage points; at 4B, it achieves the best overall performance among the compared models. Ablation studies further support the contributions of diagnostic feedback and FTR.

Our main contributions are threefold:

- We introduce ReDiS, which jointly evolves reasoning exploration and diagnosis for RSVQA. Answer-anchored diagnostic targets and correctness-gated feedback connect the two roles using answer labels without annotated reasoning trajectories.
- We propose FTR to address the loss of relative learning signals in all-incorrect Reasoner groups. Minimal outcome-level correction provides an optimization direction without prescribing visual evidence or reasoning targets.

- Experiments on LHRS-Bench, MME-RS, and CHOICE demonstrate improvements over the Qwen3.5-0.8B, 2B, and 4B backbones. ReDiS improves AA by 6.47 percentage points on average at 0.8B, while ReDiS-4B achieves the highest average OA and AA among the compared models, including larger remote sensing MLLMs.

2 Related Work

2.1 Multimodal Large Language Models in Remote Sensing

With the development of general MLLMs (Yang et al., 2025; Dai et al., 2023; Chen & Xing, 2024; Liu et al., 2023), remote sensing research is gradually shifting from task-specific models toward unified multimodal modeling frameworks. RSGPT (Hu et al., 2025) investigates the transferability of large vision-language models to remote sensing image captioning and RSVQA through high-quality remote sensing image-text data. GeoChat (Kuckreja et al., 2024) further extends these capabilities to region-level dialogue, visual grounding, and high-resolution remote sensing understanding. EarthGPT (Zhang et al., 2024) integrates multi-source data, including optical, SAR, and infrared imagery, and supports a diverse set of tasks, such as scene classification, visual question answering, grounding, and object detection, within a unified model. Subsequent studies further advance domain-specific visual understanding. LHRS-Bot (Muhtar et al., 2024) improves remote sensing understanding through large-scale image-text data and multi-level vision-language alignment, while GeoLLaVA-8K (Wang et al., 2026b) addresses visual token redundancy in ultra-high-resolution remote sensing imagery through background token pruning and anchored token selection.

Beyond advances in visual perception, recent studies have also begun to investigate explicit reasoning in remote sensing models. RSThinker (Liu et al., 2026) constructs perceptually-grounded geospatial chain-of-thought and applies supervised fine-tuning with reinforcement learning to improve answer correctness. These methods demonstrate the effectiveness of process-level supervision for complex remote sensing reasoning, but typically require additional construction and filtering of high-quality reasoning trajectories. Rather than providing predefined target reasoning trajectories, ReDiS allows the current policy to generate candidates online and shift the optimization objective from reasoning trajectory imitation to online reasoning exploration and diagnosis.

2.2 Self-Evolution and Internal Feedback Learning

Self-evolution has recently emerged as an important direction for improving reasoning models. In language models, EMPO (Zhang et al., 2026b) constructs unsupervised reward signals from semantic entropy and improves reasoning by reinforcing high-frequency semantic clusters. INTUITOR (Zhao et al., 2026) proposes Reinforcement Learning from Internal Feedback (RLIF), which replaces external correctness rewards with internal self-certainty and incorporates this signal into group-relative optimization with GRPO. This paradigm has subsequently been extended to multimodal self-evolution. Seirênes (Zhang et al., 2026a) uses adversarial self-play to co-evolve distraction generation and mathematical problem solving within a shared model. EvoLMM (Thawakar et al., 2025) instantiates a Proposer and a Solver from the same model and improves visual reasoning through online question generation and answer self-consistency using only unlabeled images.

However, internal certainty does not necessarily indicate task correctness. No Free Lunch (Zhang et al., 2025a) systematically investigates internal feedback signals, and shows that they may induce excessive policy confidence and subsequent performance degradation during late training. Self-Judge (Wu et al., 2026) further observes that majority responses can converge to an incorrect consensus and therefore introduces a frozen Judge to regulate the quality of self-consistency signals. VISE (Venkatraman et al., 2026) characterizes a related failure mode as visual under-conditioning and introduces geometric and semantic invariance rewards to directly strengthen the dependence of the model on visual content. DUEL (Qiu et al., 2026) instead constructs closely matched true and false statements through adversarial self-play between a Challenger and a Solver, requiring the Solver to perform fine-grained verification against the image. Collectively, these studies reflect a broader shift from answer-level self-consistency toward explicit visual verification. Nevertheless, most existing methods derive feedback from the same policy, a frozen auxiliary judge, or task-specific verification signals. In contrast, ReDiS moves beyond self-contained internal feedback by separating reasoning exploration from diagnosis and coupling their evolution through externally anchored supervision.

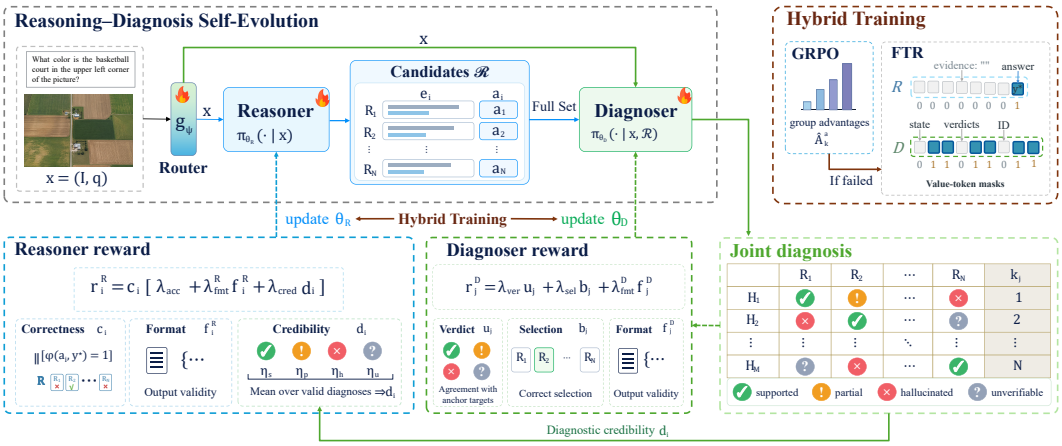

Figure 2: **Overview of ReDiS: Reasoning–Diagnosis Self-Evolution.** The Reasoner explores visual evidence–answer candidates, while the Diagnoser jointly evaluates the complete set. Answer labels anchor diagnostic learning and gate credibility feedback to correct-answer candidates. The two roles co-evolve through online sampling and separate group-relative updates. Failure-Triggered Recovery (FTR) supplies value-token correction on failed groups, leaving visual evidence and free-form reasoning without prescribed targets.

3 METHOD

3.1 DESIGN OVERVIEW

ReDiS learns directly from the model's evolving reasoning attempts while keeping self-generated feedback anchored to task correctness. A correct final answer alone does not establish whether its supporting visual claims are reliable, while agreement among multiple responses may simply reflect a shared error. We therefore decouple candidate generation from candidate evaluation. The **Reasoner** explores alternative visual evidence and answers, whereas the **Diagnoser** evaluates these candidates against the image and question. Reference answers are introduced only after rollout to anchor optimization, ensuring that diagnostic preferences cannot convert an incorrect candidate into a positive learning signal. Figure 2 illustrates this learning loop.

For each training example, the input $x = (I, q)$ consists of an image and a question, with reference answer $y^\star$. Both roles share a frozen multimodal backbone and use independent LoRA adapters (Hu et al., 2022), parameterized by θ_R and θ_D, allowing the two roles to specialize without duplicating the full model. A separately trained router g_ψ selects the corresponding role according to the input prompt. Neither role observes $y^\star$ during rollout; the reference answer is used only afterward to construct rewards and, when necessary, recovery targets. Training therefore requires neither annotated visual evidence nor teacher-generated reasoning trajectories. At inference time, only the Reasoner branch is activated for a single deterministic generation.

3.2 ONLINE EXPLORATION AND JOINT DIAGNOSIS

Exploring current reasoning attempts. Given x, the Reasoner samples multiple responses to expose alternative interpretations of the same image. Each candidate pairs a concise description of visual evidence or reasoning rationale e_i with an answer a_i:

$$R_i = (e_i, a_i) \sim \pi_{\theta_R}(\cdot \mid x), \qquad i = 1, \ldots, N. \tag{1}$$

The resulting candidates form the set $\mathcal{R}$. Because they are generated online, they reflect the current policy's visual interpretations and failure modes and therefore evolve together with the Reasoner. The explicit evidence field exposes the stated basis of each prediction to the Diagnoser without requiring a reference reasoning trajectory.

Diagnosing the complete candidate set. The Diagnoser receives the input x together with all N candidates. Observing the complete candidate set allows it to compare the candidates globally and

determine which response is the most reliable. Importantly, this global comparison does not make candidate-level judgments dependent on agreement among responses: each candidate is evaluated independently according to whether its answer and stated evidence are supported by the image and question. The Diagnoser samples M complete diagnostic responses. Each sampled diagnosis assigns a verdict $\hat{v}_{ji}$ to every candidate and returns a selection k_j, where $k_j \in \{1, \ldots, N\}$ identifies the preferred candidate and $k_j = 0$ indicates rejection of the entire set.

We use four diagnostic states: v_s for visually supported content, v_p for a supported answer accompanied by locally unreliable evidence, v_h for an evident hallucination, and v_u when the available evidence is insufficient for a reliable judgment. Candidate-level verdicts provide fine-grained feedback for individual reasoning candidates, whereas set-level selection identifies the most reliable candidate, with the option to reject the entire set. These visual judgments are learned through answer-anchored proxy targets, described next.

3.3 Hybrid Training

Reasoner Reward. After rollout, we determine candidate correctness by matching a_i to the reference answer $y^\star$ with the answer-matching function ϕ. This yields a binary indicator c_i, where unmatched or unparseable answers are treated as incorrect. We then aggregate valid diagnostic judgments into a candidate-level credibility score.

$$d_i = \frac{1}{|\mathcal{J}_{\text{valid}}|} \sum_{j \in \mathcal{J}_{\text{valid}}} s(\hat{v}_{ji}), \tag{2}$$

Here, $\mathcal{J}_{\text{valid}}$ denotes the set of diagnostic responses satisfying the candidate-binding constraints, and $d_i = 0$ when no valid diagnosis is available. The scoring function $s(\cdot)$ maps the four verdicts to the nonnegative values η_s, η_p, η_h, and η_u. We combine this credibility signal with answer correctness and structured-output validity to define the Reasoner reward:

$$r_i^R = c_i \left(\lambda_{\text{acc}} + \lambda_{\text{fmt}}^R f_i^R + \lambda_{\text{cred}} d_i \right). \tag{3}$$

Here, the nonnegative weights λ_{acc}, λ_{fmt}^R, and λ_{cred} control correctness, format, and credibility, respectively. The multiplicative correctness gate is crucial: every incorrect candidate receives zero reward regardless of how strongly the Diagnoser endorses it. Diagnostic feedback therefore refines preferences only among answer-correct candidates, providing an additional learning signal beyond the shared outcome label without allowing an incorrect consensus to receive positive reinforcement.

Diagnoser Reward. Because the Diagnoser itself evolves, its predictions require an external anchor during optimization. We derive proxy targets for the diagnostic states defined above from answer correctness: a correct answer maps to v_s, a parseable incorrect answer to v_h, and an unparseable answer to v_u. The state v_p cannot be inferred directly from the answer label, since a correct answer may still contain locally unreliable evidence. We therefore treat v_p as a softened match to v_s: when the proxy target is v_s, predicting v_p receives partial credit $\kappa \in [0, 1]$ rather than being counted as a complete mismatch. The verdict-consistency score u_j is then computed by averaging these per-candidate matching scores over all N candidates.

We reward set-level selection separately. If at least one candidate is answer-correct, the selection score b_j equals one only when the Diagnoser selects a correct candidate whose predicted verdict is either v_s or v_p. If no candidate is correct, $b_j = 1$ only when the complete set is rejected, i.e., $k_j = 0$. Otherwise, $b_j = 0$. Combining verdict consistency, selection quality, and structured-output validity yields the Diagnoser reward:

$$r_j^D = \lambda_{\text{ver}} u_j + \lambda_{\text{sel}} b_j + \lambda_{\text{fmt}}^D f_j^D, \tag{4}$$

The nonnegative weights λ_{ver}, λ_{sel}, and λ_{fmt}^D control candidate-level diagnostic consistency, set-level decision quality, and output validity, respectively. These answer-derived targets serve only as outcome anchors; they do not directly supervise evidence faithfulness or diagnostic rationales.

Self-Evolution Framework. The two roles interact through a changing online learning distribution. As the Reasoner improves, it produces new candidate sets and exposes different error patterns to the Diagnoser. As the Diagnoser evolves, its credibility estimates reshape the relative preference among successful Reasoner candidates. Repeated online sampling closes this loop: each role continually

changes the training distribution or feedback observed by the other, while reference answers remain an external anchor for task correctness. Each role performs group-relative optimization within its own rollout group. The Reasoner compares N candidates generated for the same input, whereas the Diagnoser compares M complete diagnostic responses for the same candidate set. For role $a \in \{R, D\}$, group rewards are normalized as

$$\hat{A}_k^a = \frac{r_k^a - \mu_a}{\sigma_a + \epsilon}, \qquad G_R = N, \quad G_D = M, \tag{5}$$

where G_a denotes the corresponding group size, and μ_a and σ_a are the group reward mean and standard deviation. The constant $\epsilon > 0$ ensures numerical stability. Using separate optimizers, the two role-specific adapters independently maximize the clipped GRPO objective:

$$\mathcal{J}_{\text{GRPO}}^a = \mathbb{E}\left[\frac{1}{G_a}\sum_{k=1}^{G_a}\frac{1}{T_k}\sum_{t=1}^{T_k}\min\left(\rho_{k,t}^a(\theta_a)\hat{A}_k^a, \text{clip}\left(\rho_{k,t}^a(\theta_a), 1-\varepsilon, 1+\varepsilon\right)\hat{A}_k^a\right)\right]. \tag{6}$$

Here, T_k denotes the rollout length, $\rho_{k,t}^a$ is the current-to-rollout policy probability ratio for token t under the same input and prefix, and ε is the clipping threshold. The expectation is taken over training inputs and online rollouts.

Failure-Triggered Recovery. Correctness gating prevents incorrect candidates from receiving positive reward, but it also exposes a degenerate case of group-relative optimization. When all N Reasoner candidates are incorrect, every Reasoner reward becomes zero and the corresponding group-relative advantages vanish. No sampled response then provides a useful direction for GRPO. We introduce **Failure-Triggered Recovery (FTR)** to restore a learning signal specifically for these failed groups while leaving intermediate reasoning unconstrained.

For the Reasoner, the failure set $\mathcal{F}_R$ contains inputs for which all N sampled candidates are incorrect. Recovery supervises only the reference-answer value tokens and provides neither target visual evidence nor a target reasoning trajectory. For the Diagnoser, $\mathcal{F}_D$ contains inputs for which none of the M diagnostic responses simultaneously provides valid candidate-level judgments and a valid set-level decision. Its recovery targets follow the answer-anchored verdict and selection rules defined above, with supervision restricted to diagnostic-state and candidate-selection value tokens. Free-form diagnostic rationales remain unsupervised.

We implement this restriction through a binary value-token mask $m_{x,t}^a$ over the correction target z_x^a. For a nonempty failure set, the masked cross-entropy objective is

$$\mathcal{L}_{\text{corr}}^a = -\frac{1}{|\mathcal{F}_a|}\sum_{x\in\mathcal{F}_a}\frac{\sum_t m_{x,t}^a \log \pi_{\theta_a}\left(z_{x,t}^a \mid \xi_x^a, z_{x,<t}^a\right)}{\sum_t m_{x,t}^a}, \tag{7}$$

where the conditioning input ξ_x^a is x for the Reasoner and $(x, \mathcal{R})$ for the Diagnoser. For each role, the correction step is applied after the GRPO update and is skipped when the corresponding failure set is empty. GRPO and FTR therefore remain separate optimizer steps rather than components of a weighted joint loss. Group-relative learning exploits differences among sampled responses whenever informative candidates are available, whereas FTR supplies a minimal outcome-directed signal only when those relative differences disappear.

4 Experimental Setup

Training setup. We implement ReDiS in VERL (Sheng et al., 2025) and instantiate the Reasoner and Diagnoser as two independent rank-16 LoRA (Hu et al., 2022) adapters over a frozen Qwen3.5 multimodal backbone. Training uses the GeoLLaVA-8K (Wang et al., 2026b) data release with answer labels only; no reference visual evidence or teacher-generated reasoning trajectories are used. For each question, the Reasoner samples $N = 8$ candidates and the Diagnoser produces $M = 8$ diagnostic responses. Both adapters are optimized with AdamW (Loshchilov & Hutter, 2019) using a base learning rate of 10^{-6} for up to 1,000 training steps. We set $(\lambda_{\text{acc}}, \lambda_{\text{fmt}}^R, \lambda_{\text{cred}}) = (2, 0.1, 0.25)$ and $(\lambda_{\text{ver}}, \lambda_{\text{sel}}, \lambda_{\text{fmt}}^D) = (2, 0.5, 0.5)$. The verdict scores are $(\eta_s, \eta_p, \eta_h, \eta_u) = (1, 0.5, 0, 0)$, with partial-credit coefficient $\kappa = 0.5$. Additional implementation, sampling, and optimization details are provided in Appendix A.1.

Table 1: Comparison with existing vision-language models on remote sensing benchmarks. OA and AA denote overall accuracy and average accuracy, respectively.

Method	Publication	Size	LHRS-Bench		MME-RS		CHOICE		Avg.	
			OA	AA	OA	AA	OA	AA	OA	AA
Closed-source Commercial Vision-Language Models										
GPT-4o	arXiv 2024	–	70.29	66.15	21.43	21.36	67.71	65.66	53.14	51.06
Gemini-2.5-pro	arXiv 2025	–	<u>74.35</u>	66.41	38.23	38.06	<u>75.84</u>	<u>72.48</u>	<u>62.81</u>	58.98
Claude-haiku-4.5	–	–	58.70	56.59	20.48	21.03	64.32	59.82	47.83	45.81
Open-source Vision-Language Models										
Qwen3.5-0.8B	–	0.8B	54.35	45.40	38.36	38.36	56.79	51.90	49.83	45.22
Qwen3.5-2B	–	2B	65.80	56.24	36.49	36.34	63.74	62.54	55.34	51.71
Qwen3.5-4B	–	4B	72.20	<u>69.54</u>	38.63	38.63	72.16	71.68	61.00	<u>59.95</u>
LLaVA-v1.6-vicuna	NeurIPS 2023	7B	60.72	52.87	38.98	38.87	56.29	53.32	52.00	48.35
LLaMA3-LLaVA	–	8B	66.09	55.50	**47.57**	**47.57**	65.99	63.47	59.88	55.51
DeepSeek-VL-Chat	arXiv 2024	7B	57.10	47.98	35.07	34.96	63.37	62.07	51.85	48.34
Open-source Remote Sensing Vision-Language Models										
GeoChat	CVPR 2024	7B	49.86	44.09	29.29	29.24	55.11	52.89	44.75	42.07
VHM	AAAI 2025	7B	54.78	49.42	38.47	38.38	63.13	58.35	52.13	48.72
EarthDial	CVPR 2025	4B	56.81	49.30	33.98	33.88	63.22	59.75	51.34	47.64
RemoteReasoner	AAAI 2026	7B	66.67	57.18	30.44	30.29	71.29	69.02	56.13	52.16
RSThinker	ICLR 2026	9B	65.51	58.16	38.44	38.36	62.33	56.70	55.43	51.07
Our Method										
ReDiS-0.8B	Ours	0.8B	57.83	49.87	<u>46.28</u>	<u>46.15</u>	63.13	59.05	55.75	51.69
Δ			+3.48	+4.47	+7.92	+7.79	+6.34	+7.15	+5.91	+6.47
ReDiS-2B	Ours	2B	70.14	59.97	44.44	44.59	71.84	69.65	62.14	58.07
Δ			+4.34	+3.73	+7.95	+8.25	+8.10	+7.11	+6.80	+6.36
ReDiS-4B	Ours	4B	**74.49**	**72.42**	39.54	39.35	**77.23**	**75.03**	**63.75**	**62.27**
Δ			+2.29	+2.88	+0.91	+0.72	+5.07	+3.35	+2.76	+2.32

Benchmarks and Baselines. We evaluate RSVQA on LHRS-Bench (Muhtar et al., 2024), MME-RS from MME-RealWorld (Zhang et al., 2025b), and CHOICE (An et al., 2025). We report overall accuracy (OA) and category-averaged accuracy (AA); Avg. denotes the unweighted mean across benchmarks. Table 1 compares ReDiS with three groups of MLLMs: proprietary models (GPT-4o (Hurst et al., 2024), Gemini-2.5-Pro (Comanici et al., 2025), and Claude-Haiku-4.5); open-source general-purpose models (Qwen3.5 (Team, 2026), LLaVA-v1.6-Vicuna (Liu et al., 2023), LLaMA3-LLaVA (Chen & Xing, 2024), and DeepSeek-VL-Chat (Lu et al., 2024)); and remote sensing models (GeoChat (Kuckreja et al., 2024), VHM (Pang et al., 2025), EarthDial (Soni et al., 2025), RemoteReasoner (Yao et al., 2026a), and RSThinker (Liu et al., 2026)). Qwen3.5 at 0.8B, 2B, and 4B provides the corresponding backbone references; each Δ reports the absolute gain in percentage points over the backbone at the same scale.

5 Results and Analysis

5.1 Quantitative Analysis

Table 1 shows that ReDiS consistently improves its matched Qwen3.5 backbone across all three benchmarks and model scales. The average OA gains are 5.91, 6.80, and 2.76 percentage points for the 0.8B, 2B, and 4B variants, respectively, with corresponding AA gains of 6.47, 6.36, and 2.32 points. The improvements are therefore consistent under both question-level and category-averaged evaluation. The 2B model exhibits the largest average OA gain, driven in particular by improvements of 7.95 points on MME-RS and 8.10 points on CHOICE.

Comparison with existing models. ReDiS-4B achieves the highest average OA (63.75%) and AA (62.27%) among the models reported in Table 1. It also attains the strongest OA/AA results on LHRS-Bench and CHOICE, reaching 74.49%/72.42% and 77.23%/75.03%, respectively. Compared with Gemini-2.5-Pro, the strongest proprietary baseline by average OA, ReDiS-4B improves average

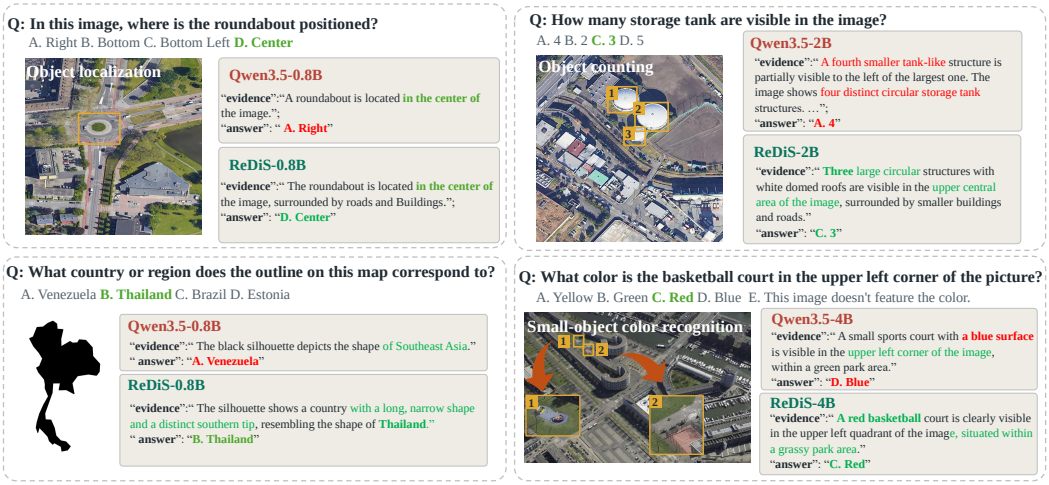

Figure 3: Selected response excerpts from ReDiS and the corresponding Qwen3.5 backbones. The examples cover object localization (top left), map recognition (bottom left), object counting (top right), and small-object color recognition (bottom right). For ease of interpretation, key visual evidence is highlighted with yellow boxes.

OA by 0.94 and AA by 3.29 percentage points. ReDiS-2B likewise exceeds all listed remote sensing baselines in both average metrics, outperforming RemoteReasoner by 6.01 OA and 5.91 AA points.

Performance across model scales. ReDiS-0.8B reaches 55.75% average OA and 51.69% average AA, closely matching Qwen3.5-2B at 55.34% and 51.71% despite using a smaller backbone. Performance generally increases with model scale, but the trend is not uniform across benchmarks. On MME-RS, ReDiS-0.8B achieves 46.28% OA and 46.15% AA, outperforming the larger ReDiS variants and ranking second among the models in Table 1, behind LLaMA3-LLaVA-8B at 47.57% on both metrics. These results indicate that the gains over matched backbones are consistent, while the optimal model scale remains task dependent. Fine-grained analyses are provided in Appendix A.2.1.

5.2 Qualitative Analysis

Figure 3 compares ReDiS with the corresponding Qwen3.5 backbones on four representative visual questions. In the localization example, Qwen3.5-0.8B correctly describes the roundabout as being at the image center but outputs the inconsistent option "Right." ReDiS instead aligns its final prediction with the stated visual evidence and selects "Center." In the map-recognition example, the backbone provides only a broad regional description and selects Venezuela, whereas ReDiS identifies Thailand and relates its prediction to the elongated outline and distinctive southern extension. The counting and color examples illustrate corrections of more localized visual errors. Qwen3.5-2B introduces a fourth tank-like structure and consequently overcounts the target objects, whereas ReDiS-2B identifies the three annotated storage tanks. Similarly, Qwen3.5-4B describes the queried basketball court as blue, while ReDiS-4B identifies the highlighted target as red. Together, these cases illustrate improvements in evidence–answer consistency, object counting, local attribute recognition, and map-shape interpretation. More qualitative comparisons are provided in Appendix A.4.

5.3 Ablation Studies

Table 2 evaluates the contributions of the Diagnoser and FTR on Qwen3.5-0.8B. The complete ReDiS model achieves the highest OA and AA across all three benchmarks. Removing the Diagnoser reduces average OA/AA from 55.75%/51.69% to 54.57%/50.24%, with the largest degradation on LHRS-Bench (2.61 OA and 3.29 AA points). The smaller differences on MME-RS and CHOICE indicate that the benefit of diagnostic feedback varies across benchmarks.

Table 2: Component ablations on Qwen3.5-0.8B. OA and AA are in percent; Δ denotes gains over the unadapted backbone.

Models	LHRS-Bench		MME-RS		CHOICE		Avg.	
	OA	AA	OA	AA	OA	AA	OA	AA
Qwen3.5-0.8B	54.35	45.40	38.36	38.36	56.79	51.90	49.83	45.22
w/o FTR (step=100)	56.81	48.97	45.40	45.27	59.29	56.80	53.83	50.35
Δ	(+2.46)	(+3.57)	(+7.04)	(+6.91)	(+2.50)	(+4.90)	(+4.00)	(+5.13)
w/o Diagnoser	55.22	46.58	45.63	45.49	62.85	58.65	54.57	50.24
Δ	(+0.87)	(+1.18)	(+7.27)	(+7.13)	(+6.06)	(+6.75)	(+4.73)	(+5.02)
ReDiS-0.8B	**57.83**	**49.87**	**46.28**	**46.15**	**63.13**	**59.05**	**55.75**	**51.69**
Δ	(+3.48)	(+4.47)	(+7.92)	(+7.79)	(+6.34)	(+7.15)	(+5.91)	(+6.47)

The w/o FTR variant disables only failure-triggered correction while retaining the remaining training configuration. Its step-100 checkpoint reaches 53.83% average OA and 50.35% average AA, whereas the full model reaches 55.75% and 51.69%, respectively. The largest improvement from FTR occurs on CHOICE, where ReDiS gains 3.84 OA and 2.25 AA points. These results indicate that the Diagnoser and FTR provide complementary benefits: diagnostic feedback refines learning when informative candidates are available, whereas FTR supplies a recovery signal when group-relative supervision collapses.

Figure 4 further illustrates the effect of FTR on training dynamics. Without FTR, performance improves rapidly from the 51.90% unadapted baseline to 56.80% near step 100, but subsequently becomes unstable and gradually declines. This behavior is consistent with the failure mode identified in Sec. 3.4: when all sampled candidates are incorrect, correctness-gated rewards become identical and provide no group-relative learning signal. With FTR enabled, such failed groups receive a minimal outcome-level correction while intermediate reasoning remains unsupervised. ReDiS therefore continues to improve beyond the early stage, reaching 59.05% near step 700 and remaining comparatively stable thereafter. This corresponds to a 7.15-point improvement over the unadapted baseline and a 2.25-point gain over the best checkpoint of the w/o FTR variant. The results support FTR as a complementary recovery mechanism that restores an optimization direction when relative feedback is unavailable.

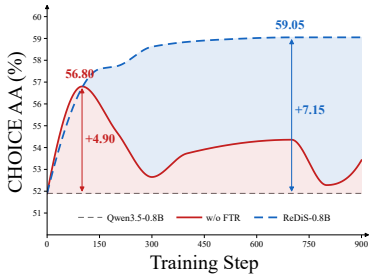

Figure 4: Training dynamics of Qwen3.5-0.8B w/ and w/o FTR.

6 Discussion and Future Research

ReDiS enables online exploration without prescribed reasoning trajectories, but still relies on supplied questions and answer labels. These labels anchor rewards and provide recovery targets when exploration fails. Moreover, answer-anchored diagnostic targets are proxies: answer correctness does not establish the faithfulness of every intermediate claim. The current RSVQA results therefore support improved answer prediction, while evidence-level diagnosis requires further evaluation. Recent progress in self-evolving visual question generation (He et al., 2026) motivates extending ReDiS to post-training on unlabeled images without supplied questions or answer labels. Such a system could autonomously select regions of interest, formulate questions, generate evidence–answer candidates, and diagnose their reliability. Diagnosis could then guide subsequent inspection and question generation. The central challenge is to replace ReDiS's label-dependent correctness and recovery signals with reliable visual feedback. Visual invariance constraints and region-level semantic verification offer possible starting points (Venkatraman et al., 2026; Yao et al., 2026b), but their suitability for joint candidate diagnosis and failure recovery remains to be established. In particular, the system would need to identify when self-generated feedback is unreliable and determine how to recover without an externally verified target.

7 Conclusion

We presented ReDiS, a framework for improving remote sensing multimodal reasoning through collaborative Reasoning–Diagnosis Self-evolution. Answer-anchored diagnosis and correctness-gated rewards guide learning from online candidates, while FTR supplies minimal value-token supervision when exploration provides no usable feedback. This design requires no reference reasoning trajectories and retains single-pass Reasoner inference. Across LHRS-Bench, MME-RS, and CHOICE, ReDiS improves accuracy over the corresponding Qwen3.5 backbones. These results support using outcome supervision to guide joint reasoning and diagnostic learning while leaving intermediate reasoning self-explored.

AI USE STATEMENT

We used generative AI tools to assist manuscript drafting, language refinement, and LaTeX editing. These tools were not used to create additional training data or annotations for this study. The authors take responsibility for the final manuscript, including its technical claims, experimental results, and references.

ETHICS STATEMENT

This work studies remote sensing image understanding using existing datasets. Potential benefits include environmental monitoring and disaster assessment, while risks include incorrect geographic interpretations, biases inherited from training data, and privacy-sensitive inferences about locations or infrastructure. Answer accuracy does not guarantee faithful visual evidence; consequential applications require independent validation and human oversight. Use and redistribution of third-party imagery and annotations remain subject to their original terms.

REPRODUCIBILITY STATEMENT

The Method section specifies the rewards, group-relative objectives, and FTR. Appendix A.1 documents the ReDiS hybrid configuration, data filtering and splitting, optimization schedule, checkpoint selection, and RSVQA evaluation protocol. We plan to release training and evaluation code, configuration files, and data preparation scripts with the camera-ready version. Third-party datasets will be referenced through their original releases.

REFERENCES

Xiao An, Jiaxing Sun, Zihan Gui, and Wei He. CHOICE: Benchmarking the remote sensing capabilities of large vision-language models. In *Advances in Neural Information Processing Systems*, volume 38, 2025.

Christel Chappuis, Valérie Zermatten, Sylvain Lobry, Bertrand Le Saux, and Devis Tuia. Prompt–rsvqa: Prompting visual context to a language model for remote sensing visual question answering. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 1371–1380. IEEE, 2022.

Lin Chen and Long Xing. Open-llava-next: An open-source implementation of llava-next series for facilitating the large multi-modal model community. *GitHub-xiaoachen98/Open-LLaVA-NeXT: Anopen-sourceimplementationfortrainingLLaVA-NeXT*, 2024.

Qimin Cheng, Haiyan Huang, Yuan Xu, Yuzhuo Zhou, Huanying Li, and Zhongyuan Wang. Nwpu-captions dataset and mlca-net for remote sensing image captioning. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–19, 2022.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*, 2025.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale N Fung, and Steven Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning. *Advances in neural information processing systems*, 36:49250–49267, 2023.

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.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022.

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

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.

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.

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, 2026.

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.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019.

Haoyu Lu, Wen Liu, Bo Zhang, Bingxuan Wang, Kai Dong, Bo Liu, Jingxiang Sun, Tongzheng Ren, Zhuoshu Li, Hao Yang, Yaofeng Sun, Chengqi Deng, Hanwei Xu, Zhenda Xie, and Chong Ruan. Deepseek-vl: Towards real-world vision-language understanding. *arXiv preprint arXiv:2403.05525*, 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.

Guangming Sheng, Chi Zhang, Zilingfeng Ye, Xibin Wu, Wang Zhang, Ru Zhang, Yanghua Peng, Haibin Lin, and Chuan Wu. Hybridflow: A flexible and efficient rlhf framework. In *Proceedings of the Twentieth European Conference on Computer Systems*, pp. 1279–1297, 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 *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14303–14313. IEEE, 2025.

Qwen Team. Qwen3.5: Accelerating productivity with native multimodal agents, February 2026. URL `https://qwen.ai/blog?id=qwen3.5`.

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.

Shravan Venkatraman, Ritesh Thawkar, Omkar Thawakar, Rao Muhammad Anwer, Hisham Cholakkal, Salman Khan, and Fahad Shahbaz Khan. Paying more attention to visual tokens in self-evolving large multimodal models. In *European Conference on Computer Vision*, pp. 41–59. Springer, 2026.

Changzhi Wang, Guowei Zheng, Renyue Zhang, Yuhan Zhu, Fengjun Zhou, and Xing Liu. Dppf: Dual-path pre-fusion with semantic-guided encoding for remote sensing image captioning. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 2026a.

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, 2026b.

Junjue Wang, Zhuo Zheng, Zihang Chen, Ailong Ma, and Yanfei Zhong. Earthvqa: Towards queryable earth via relational reasoning-based remote sensing visual question answering. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, pp. 5481–5489, 2024.

Zhengxian Wu, Kai Shi, Chuanrui Zhang, Zirui Liao, Jun Yang, Ni Yang, Qiuying Peng, Luyuan Zhang, Hangrui Xu, Tianhuang Su, et al. When models judge themselves: Unsupervised self-evolution for multimodal reasoning. *arXiv preprint arXiv:2603.21289*, 2026.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

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, 2026a.

Liang Yao, Fan Liu, Shengxiang Xu, Chuanyi Zhang, Rui Min, Shimin Di, and Yuhui Zheng. Remotezero: Geospatial reasoning with zero labels. *arXiv preprint arXiv:2605.04451*, 2026b.

Qian Yu, Peng Gong, Nick Clinton, Greg Biging, Maggi Kelly, and Dave Schirokauer. Object-based detailed vegetation classification with airborne high spatial resolution remote sensing imagery. *Photogrammetric Engineering & Remote Sensing*, 72(7):799–811, 2006.

Chi Zhang, Haibo Qiu, Qiming Zhang, Yufei Xu, Xinbo Gao, and Jing Zhang. Seirênes: Adversarial self-play with evolving distractions for LLM reasoning. *arXiv preprint arXiv:2605.11636*, 2026a.

Qingyang Zhang, Haitao Wu, Changqing Zhang, Peilin Zhao, and Yatao Bian. Right question is already half the answer: Fully unsupervised llm reasoning incentivization. *Advances in neural information processing systems*, 38:67345–67372, 2026b.

Wei Zhang, Miaoxin Cai, Tong Zhang, Yin Zhuang, and Xuerui Mao. Earthgpt: A universal multimodal large language model for multisensor image comprehension in remote sensing domain. *IEEE Transactions on Geoscience and Remote Sensing*, 62:1–20, 2024.

Yanzhi Zhang, Zhaoxi Zhang, Haoxiang Guan, Yilin Cheng, Yitong Duan, Chen Wang, Yue Wang, Shuxin Zheng, and Jiyan He. No free lunch: Rethinking internal feedback for llm reasoning. *arXiv preprint arXiv:2506.17219*, 2025a.

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, 2025b.

Xuandong Zhao, Zhewei Kang, Aosong Feng, Sergey Levine, and Dawn Song. Learning to reason without external rewards. In *International Conference on Learning Representations*, volume 2026, pp. 2548–2581, 2026.

A Appendix

A.1 Experimental Details

A.1.1 Hybrid training configuration

Implementation. We instantiate ReDiS with Qwen3.5 and implement hybrid training in VERL using Ray and Fully Sharded Data Parallel (FSDP). The multimodal backbone remains frozen; the Reasoner and Diagnoser use independent LoRA adapters with rank 16, LoRA alpha 32, and zero dropout. LoRA is applied to attention, linear-attention, and feed-forward projections. We use BF16 mixed-precision computation and gradient checkpointing.

Data and preprocessing. Training data are drawn from the GeoLLaVA-8K data release (Wang et al., 2026b). We retain single-image, four-option questions with one unambiguous answer label, yielding 79,666 valid examples after conversion. No reference evidence or reasoning trajectory is used as a training target. We reserve approximately 2% of image-path groups with a fixed split seed of 20260907 and exclude all questions associated with held-out images from training. The resulting training set contains 77,659 questions; checkpoint selection uses a fixed subset of 256 held-out questions from 59 images. Images are processed with a pixel budget of 262,144–1,310,720 pixels rather than a fixed square resolution.

Online sampling. Each training step samples $N = 8$ Reasoner candidates and $M = 8$ diagnoses per question, with four source questions per GPU. Both roles use temperature 1.0, top-$p = 0.95$, and top-$k = 20$. The Reasoner and Diagnoser have input limits of 4,096 and 12,288 tokens and output limits of 128 and 768 tokens, respectively. Native thinking mode is disabled; the Reasoner produces a short evidence field followed by an answer. Diagnostic prompts that exceed the input budget are skipped as complete candidate sets rather than truncated.

Optimization and checkpoint selection. The two adapters use separate AdamW optimizers with a base learning rate of 10^{-6}, zero weight decay, gradient clipping at 1.0, and a GRPO clipping threshold of 0.2. We update the Diagnoser before the Reasoner. Learning rates warm up for 10 steps and then follow cosine decay to 10% of the base rate at step 500, remaining at that floor thereafter. For each role, the GRPO learning rate is further scaled by the fraction of groups with nonzero advantages. FTR uses a separate optimizer step with learning-rate multipliers of 0.2 and 0.3 for the Reasoner and Diagnoser, respectively; these are not joint-loss weights. We use neither an additional reference-policy KL penalty nor entropy regularization.

A.1.2 RSVQA evaluation and baselines

Benchmark scope and scoring. LHRS-Bench is abbreviated as LHRS, and the MME-RealWorld results are denoted MME-RS, as in Table 1. For CHOICE, we evaluate only text-answer multiple-choice questions: coordinate targets and records requiring bounding-box or segmentation outputs are excluded. Scores therefore refer to this subset rather than the full heterogeneous release. Each input contains the image, question, and available options; reference answers are accessed only for scoring. Evaluation uses one deterministic Reasoner generation of at most 128 tokens, without a Diagnoser pass, voting, or best-of-N selection. We extract the predicted option and score exact agreement with the reference, counting unparseable answers and inference failures as incorrect.

OA averages correctness over all questions. AA is the unweighted mean of category accuracies, using the primary fine-grained category field (falling back to subtask or task metadata when necessary); a question with multiple category labels contributes to each assigned category. Both metrics are reported in percent, and each Avg. column averages its metric equally across the three benchmarks. These metrics assess final answers and do not directly measure evidence faithfulness.

Comparison models. The proprietary baselines are GPT-4o, Gemini-2.5-Pro, and Claude-Haiku-4.5. The general-purpose open-source baselines are Qwen3.5 at 0.8B, 2B, and 4B, LLaVA-v1.6-Vicuna-7B, LLaMA3-LLaVA-8B, and DeepSeek-VL-Chat-7B. The remote sensing baselines are GeoChat-7B, VHM-7B, EarthDial-4B, RemoteReasoner-7B, and RSThinker-9B. Each ReDiS variant is compared with the Qwen3.5 backbone at the same scale. For the 0.8B configuration, the un-

modified checkpoint is loaded without adapters or the router and evaluated with the same Reasoner prompt, answer parser, deterministic decoding, and scoring procedure. Cross-model comparisons reflect differences in pretraining and adaptation and do not isolate individual ReDiS components.

A.2 Additional Quantitative Analysis

A.2.1 Fine-Grained Category Analysis on RSVQA Benchmarks

We conduct a category-wise comparison of the three ReDiS variants against five remote sensing MLLMs. The fine-grained results reveal task-specific and scale-dependent performance differences that are not apparent from aggregate accuracy alone. Unless otherwise specified, all reported improvements are absolute percentage-point differences, and rankings are computed only among the models included in each table.

MME-RS. Table 3 shows complementary strengths across model scales. ReDiS-4B leads Color at 52.11%, ReDiS-0.8B leads Count at 29.93%, and ReDiS-2B leads Position at 62.05%, exceeding the strongest non-ReDiS baseline in each category by 12.75, 1.87, and 12.33 points, respectively. Counting remains difficult: accuracy falls from 29.93% at 0.8B to 15.42% at 4B, while Position peaks at 2B. These differences explain why the largest ReDiS variant does not achieve the highest overall MME-RS accuracy despite its strong color recognition.

Table 3: Fine-grained accuracy (%) on MME-RS. AA is the unweighted mean of the displayed category accuracies. Bold and underline denote the best and second-best results, respectively.

Method	Color	Count	Position	AA
EarthDial	35.06	22.51	44.07	33.88
GeoChat	33.71	22.92	31.11	29.25
RSThinker	39.36	<u>28.06</u>	47.65	38.36
RemoteReasoner	32.35	11.42	47.10	30.29
VHM	38.09	27.32	49.72	38.38
ReDiS-0.8B	<u>51.08</u>	**29.93**	<u>57.44</u>	**46.15**
ReDiS-2B	47.73	23.98	**62.05**	<u>44.59</u>
ReDiS-4B	**52.11**	15.42	50.52	39.35

LHRS-Bench. In Table 4, at least one ReDiS variant attains the best or joint-best accuracy in every category. ReDiS-4B leads seven of the eleven categories, including ties, with particularly large margins over the strongest non-ReDiS baselines in Distance ($+18.18$), Modality ($+17.39$), and Resolution ($+42.85$). ReDiS-2B instead leads Color and Shape and shares the best scores in Area, Quantity, and Reasoning. The 85.71% Resolution accuracy of ReDiS-4B contrasts with 9.52% for both smaller variants, while its Quantity accuracy is lower. Thus, higher aggregate accuracy coexists with substantial variation in individual capabilities.

Table 4: Fine-grained accuracy (%) on LHRS. AA is the unweighted mean of the displayed category accuracies. Bold and underline denote the best and second-best results, respectively.

Method	Area	Color	Distance	Identity	Location	Modality	Orientation	Quantity	Reasoning	Resolution	Shape	AA
EarthDial	58.67	59.29	45.45	58.99	66.18	21.74	23.08	40.88	63.04	<u>42.86</u>	62.16	49.30
GeoChat	46.67	37.17	40.91	51.26	50.49	21.74	28.21	39.42	69.57	<u>42.86</u>	56.76	44.09
RSThinker	61.33	65.49	<u>50.00</u>	67.67	70.59	<u>60.87</u>	**48.72**	46.72	<u>80.43</u>	9.52	78.38	58.16
RemoteReasoner	**68.00**	63.72	45.45	69.24	73.04	30.43	**48.72**	45.26	76.09	33.33	75.68	57.18
VHM	64.00	52.21	36.36	56.31	54.90	43.48	35.90	47.45	73.91	14.29	64.86	49.42
ReDiS-0.8B	52.00	71.68	45.45	60.41	65.69	30.43	25.64	**51.09**	71.74	9.52	64.86	49.87
ReDiS-2B	**68.00**	**80.53**	<u>50.00</u>	<u>73.03</u>	<u>76.96</u>	30.43	<u>46.15</u>	**51.09**	**84.78**	9.52	**89.19**	<u>59.97</u>
ReDiS-4B	<u>66.67</u>	<u>75.22</u>	**68.18**	**73.82**	**79.90**	**78.26**	**48.72**	<u>48.91</u>	**84.78**	**85.71**	<u>86.49</u>	**72.42**

CHOICE perception. ReDiS-4B achieves the best accuracy on six of the fourteen evaluated perception tasks in Table 5, including all three Cross-instance Discernment tasks. Its margins over

the strongest non-ReDiS baselines are 10.29 points in Attribute Comparison, 19.00 in Spatial Relationship, and 13.50 in Change Detection. It also leads Image Quality, Image Modality, and Attribute Recognition. However, Hallucination Detection remains a pronounced weakness: ReDiS-4B reaches 40.60%, compared with VHM's 92.80%, despite improving over the smaller ReDiS variants. This benchmark task measures the final Reasoner's answers; it does not directly evaluate the Diagnoser's intermediate judgments.

Table 5: Fine-grained perception accuracy (%) on CHOICE. Level-3 tasks are grouped by their Level-2 categories. AA is the unweighted mean of the displayed category accuracies. Bold and underline denote the best and second-best results, respectively. **ILC:** Image-level Comprehension; SC: Scene Classification, MR: Map Recognition, IC: Image Caption, IQ: Image Quality, IM: Image Modality. **SII:** Single-instance Identification; LR: Landmark Recognition, OP: Object Presence, OL: Object Localization, OC: Object Counting, AR: Attribute Recognition, HD: Hallucination Detection. **CID:** Cross-instance Discernment; AC: Attribute Comparison, SR: Spatial Relationship, CD: Change Detection.

Method	ILC					SII						CID			AA
	SC	MR	IC	IQ	IM	LR	OP	OL	OC	AR	HD	AC	SR	CD	
EarthDial	**97.50**	29.30	88.50	34.50	59.62	81.00	**97.40**	88.40	48.00	50.40	64.80	74.00	34.67	62.50	65.04
GeoChat	92.75	47.77	73.50	25.75	33.88	79.00	93.00	81.00	26.60	51.60	33.20	71.14	29.67	47.00	56.13
RSThinker	97.20	22.93	78.00	38.50	69.88	71.00	94.80	77.80	**67.20**	49.60	7.20	75.71	39.33	57.50	60.48
RemoteReasoner	93.70	**53.50**	**97.50**	52.25	67.25	**88.00**	92.40	75.00	50.40	63.00	77.20	76.57	55.67	62.00	71.75
VHM	96.75	33.12	82.00	39.75	63.88	77.00	93.40	**89.80**	35.00	55.20	**92.80**	46.57	37.33	59.50	64.44
ReDiS-0.8B	91.35	29.30	91.00	50.25	53.25	60.00	92.20	65.80	54.80	60.60	18.80	74.57	43.67	52.00	59.83
ReDiS-2B	92.70	48.41	93.00	56.13	70.63	78.00	93.00	80.40	54.80	63.40	27.40	84.86	62.33	64.50	69.25
ReDiS-4B	95.40	50.96	96.50	**64.50**	**78.38**	87.00	95.00	85.40	64.80	**64.20**	40.60	**86.86**	**74.67**	**76.00**	**75.73**

CHOICE reasoning. Table 6 shows that ReDiS-4B leads four of the seven reasoning tasks: Time Property, Physical Property, Resource Assessment, and Situation Inference. The first two improve over the strongest non-ReDiS baselines by 16.50 and 7.67 points, respectively. ReDiS-2B ties RemoteReasoner at 89.00% on Disaster Discrimination. Remaining gaps include Environmental Assessment, where ReDiS-4B trails EarthDial by 4.50 points, and Geospatial Determination, where it trails RemoteReasoner by 6.00 points. These results characterize the final models' category profiles; without category-level scores for the unadapted Qwen3.5 backbones, they do not isolate the contribution of self-evolution to each task.

Table 6: Fine-grained reasoning accuracy (%) on CHOICE. Level-3 tasks are grouped by their Level-2 categories. AA is the unweighted mean of the displayed category accuracies. Bold and underline denote the best and second-best results, respectively. **AttR:** Attribute Reasoning; TP: Time Property, PP: Physical Property. **AssR:** Assessment Reasoning; EA: Environmental Assessment, RA: Resource Assessment. **CSR:** Common Sense Reasoning; DD: Disaster Discrimination, GD: Geospatial Determination, SI: Situation Inference.

Method	AttR		AssR		CSR			AA
	TP	PP	EA	RA	DD	GD	SI	
EarthDial	39.50	50.67	**47.50**	37.33	86.50	64.50	78.00	57.71
GeoChat	21.50	48.00	40.50	36.33	75.00	82.50	74.00	53.98
RSThinker	46.50	53.33	40.00	64.17	67.50	58.50	70.67	57.24
RemoteReasoner	44.50	62.33	43.00	54.67	**89.00**	**89.00**	81.33	66.26
VHM	41.00	41.33	34.00	37.50	74.00	76.00	76.67	54.36
ReDiS-0.8B	41.50	56.33	26.50	51.50	68.50	54.00	60.00	51.19
ReDiS-2B	52.00	51.67	40.50	66.67	**89.00**	80.50	78.67	65.57
ReDiS-4B	**63.00**	**70.00**	43.00	**67.00**	88.00	83.00	**86.00**	**71.43**

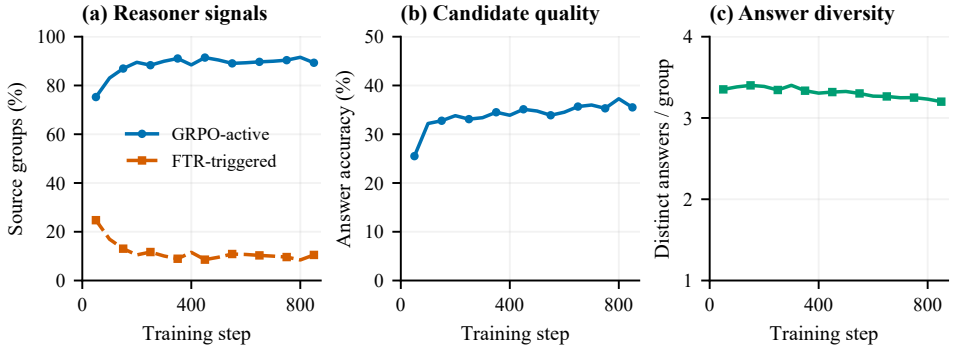

Figure 5: Online Reasoner training dynamics of ReDiS-0.8B. Each point summarizes a non-overlapping window of 50 steps (1,600 source-question presentations), with eight candidates per question. (a) Fractions of source groups with nonzero GRPO advantages and groups routed to FTR because all candidates are incorrect. (b) Candidate-level answer accuracy. (c) Mean number of distinct candidate answers.

A.3 Online Reasoner training dynamics of ReDiS-0.8B

Figure 5 characterizes how the Reasoner's online candidate distribution evolves during training. From steps 1–50 to 801–850, the fraction of all-incorrect groups decreases from 24.75% to 10.50%, while the fraction with nonzero GRPO advantages increases from 75.25% to 89.31%. Over the same period, candidate-level answer accuracy rises from 25.51% to 35.51%, whereas the mean number of distinct answers remains relatively stable. Across 27,200 source-question presentations through step 850, FTR is activated for 3,137 all-incorrect groups (11.53%). These groups would otherwise provide zero correctness-gated group-relative advantage. The statistic therefore quantifies how often the recovery path is invoked, rather than its relative contribution to the total optimization signal. Correction steps occur in 823 of 850 Reasoner steps and all 850 Diagnoser steps because a step contains multiple source groups and requires only one failed group to trigger a correction update.

A.4 Additional Qualitative Analysis

Figures 6–17 compare ReDiS with the corresponding Qwen3.5 backbones at 0.8B, 2B, and 4B across 24 selected cases. Each example presents the image, question, and evidence and answer excerpts; red answer options indicate the references.

Fine-grained visual discrimination. Figures 6, 11–14, and 17 illustrate corrections in local attributes, spatial localization, and counting. ReDiS distinguishes vehicle and roof colors, locates the specified ship and fountain, and separates individual aircraft and circular structures. It also rejects an unsupported airplane prediction. For snow and cloud coverage, ReDiS relates covered regions to the surrounding terrain and selects the reference percentage categories. These cases illustrate the importance of identifying the queried target and its relevant attributes rather than relying on a general scene description.

Consistency between evidence and answers. Figures 7, 8, and 15 show that describing relevant visual content does not always lead to the correct option. Qwen3.5 describes the red vehicle as left and below but selects top left, and identifies soccer fields but answers negatively. The junkyard example further separates perception from question interpretation: both models describe similar visual features, yet only ReDiS correctly handles the negation. ReDiS aligns the selected options with the relevant observations in these cases, while also correcting the parking-lot occupancy description.

Scene and image interpretation. Figures 9, 10, and 16 extend the comparison to modality, corruption, season, temporal change, disaster type, and map shape. ReDiS selects the displayed reference answers, and its responses cite cues such as vegetation, a newly visible football field, and structural collapse. Correct answers can still accompany incomplete explanations, as the residual RGB/false-color ambiguity illustrates. Together, these selected comparisons complement the quantitative results by characterizing specific errors and corrections, without estimating their prevalence.

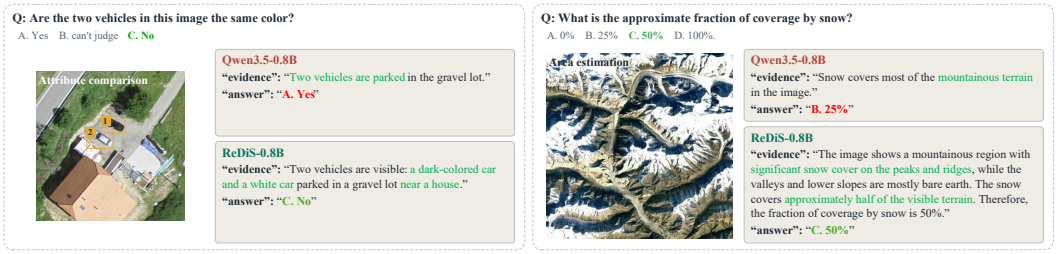

Figure 6: Attribute comparison and area estimation at 0.8B. Left: ReDiS distinguishes the dark and white vehicles, whereas Qwen3.5 selects matching colors. Right: ReDiS selects the reference 50% snow-cover category rather than the backbone's 25%.

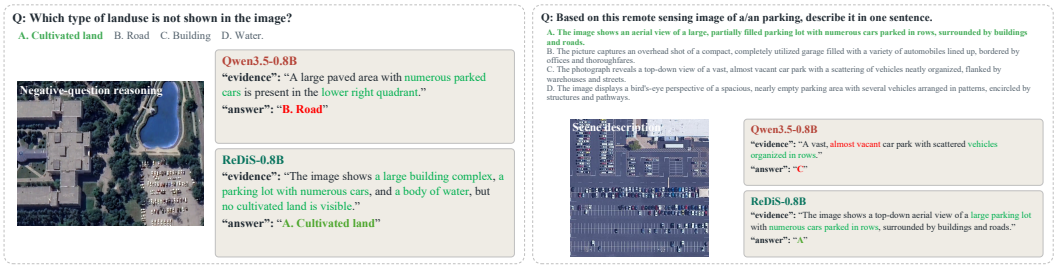

Figure 7: Land-use reasoning and scene description at 0.8B. Left: ReDiS identifies cultivated land as absent, whereas Qwen3.5 selects road. Right: ReDiS selects a partially filled parking lot with cars in rows, correcting the backbone's almost-vacant description.

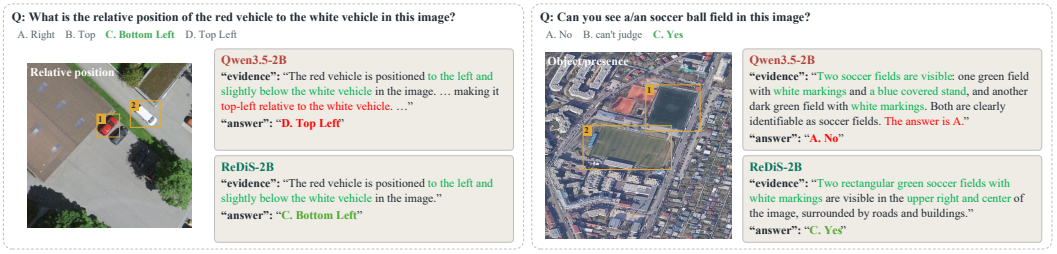

Figure 8: Evidence–answer consistency at 2B. Left: Qwen3.5 describes the red vehicle as left and below but selects top left; ReDiS selects bottom left. Right: both models describe soccer fields, but only ReDiS selects the affirmative answer.

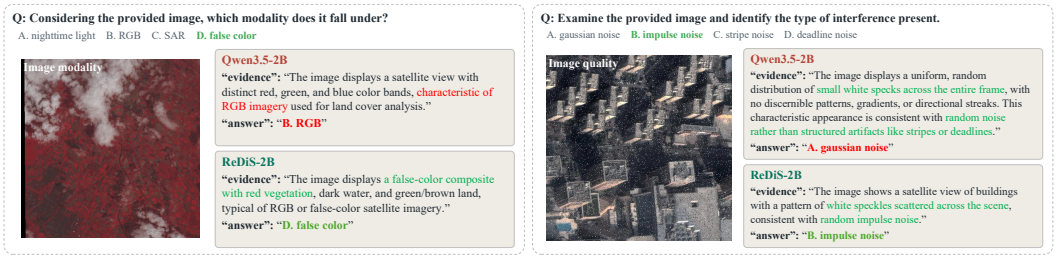

Figure 9: Image-modality and corruption recognition at 2B. Left: ReDiS selects the false-color reference rather than RGB, although its explanation retains modality ambiguity. Right: ReDiS identifies the reference impulse-noise category, whereas Qwen3.5 selects Gaussian noise.

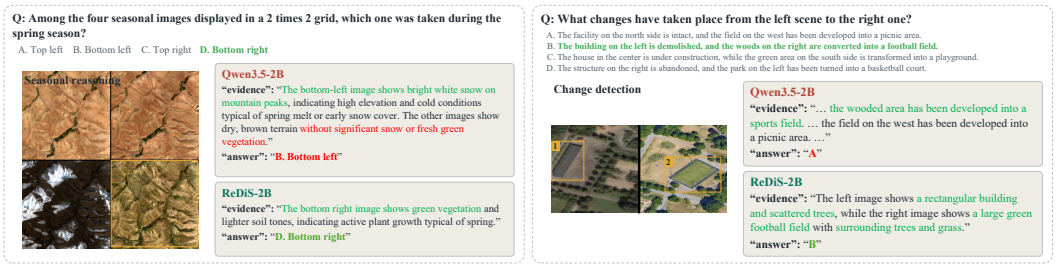

Figure 10: Seasonal interpretation and change detection at 2B. Left: ReDiS selects the bottom-right spring scene rather than the snowy bottom-left scene. Right: ReDiS selects the building-removal and football-field option; Qwen3.5 mentions a sports field but chooses the picnic-area option.

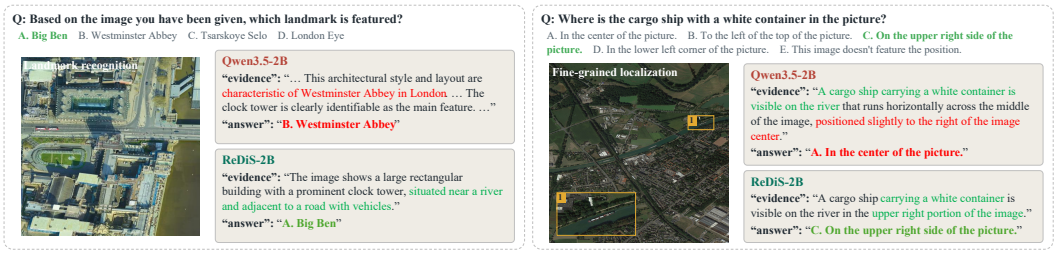

Figure 11: Landmark recognition and object localization at 2B. Left: ReDiS identifies Big Ben, whereas Qwen3.5 selects Westminster Abbey despite mentioning a clock tower. Right: ReDiS locates the specified cargo ship in the upper-right region rather than the center.



Figure 12: Small-target recognition at 2B. Left: ReDiS locates the fountain in the lower-left water area, while Qwen3.5 reports it as absent. Right: ReDiS selects yellow for the ship-deck ring rather than the backbone's green, matching the reference.

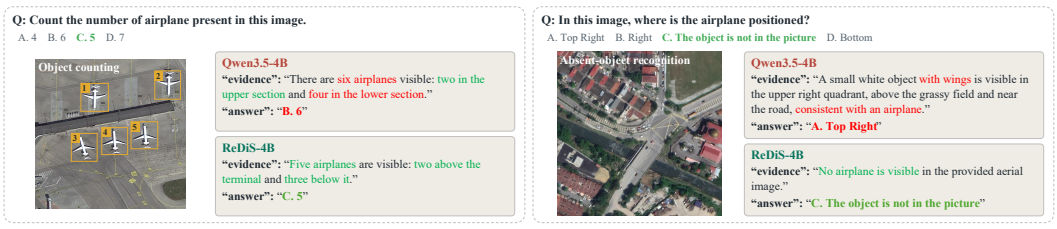

Figure 13: Object counting and absence detection at 4B. Left: ReDiS counts five aircraft, matching the annotated instances, while Qwen3.5 reports six. Right: ReDiS reports no airplane; the backbone incorrectly places one in the upper-right region.

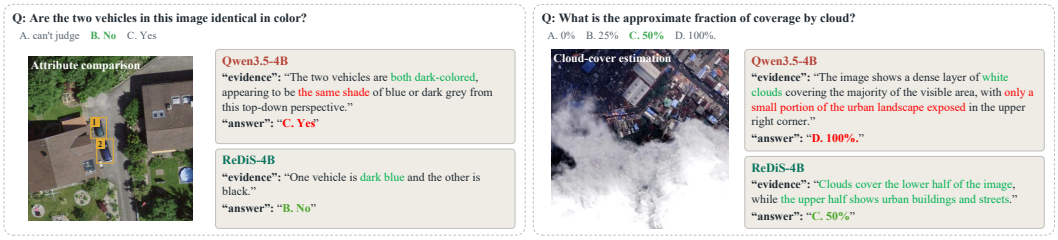

Figure 14: Color comparison and coverage estimation at 4B. Left: ReDiS distinguishes the two vehicle colors, while Qwen3.5 treats them as identical. Right: ReDiS selects the reference 50% cloud-cover category instead of the backbone's 100%.

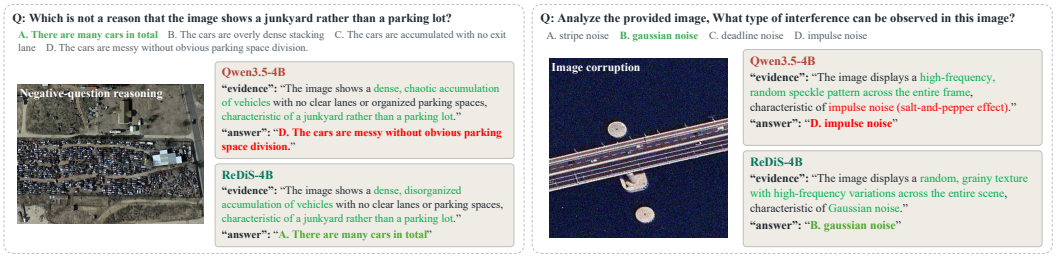

Figure 15: Question interpretation and corruption recognition at 4B. Left: despite similar junkyard descriptions, ReDiS correctly excludes vehicle count as a distinguishing reason. Right: ReDiS selects the reference Gaussian-noise label, whereas Qwen3.5 selects impulse noise.

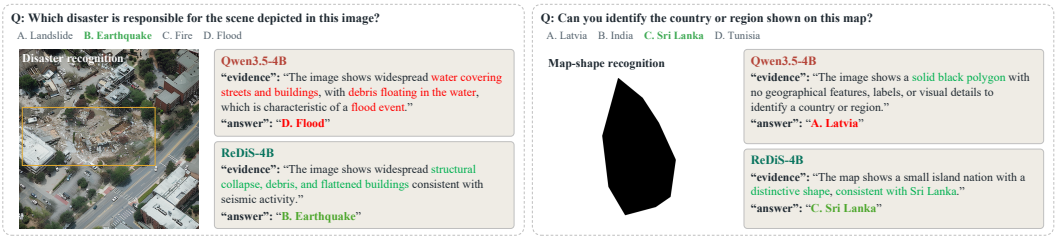

Figure 16: Disaster and map recognition at 4B. Left: ReDiS describes structural collapse and selects the earthquake reference label; Qwen3.5 selects flood. Right: ReDiS identifies Sri Lanka from its outline, while the backbone selects Latvia.

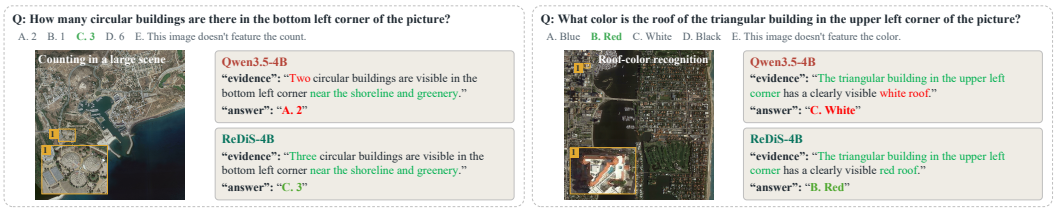

Figure 17: Regional counting and roof-color recognition at 4B. Left: ReDiS reports three circular structures in the specified lower-left region rather than two. Right: ReDiS selects red for the indicated triangular roof, correcting Qwen3.5's white prediction.

A.5 PROMPT TEMPLATES

A.5.1 TRAINING PROMPTS

We reproduce the hybrid training prompts, with line breaks adjusted for readability. Braced placeholders are filled before generation, and images are supplied as multimodal inputs. Reference answers are excluded from rollout inputs. Native thinking mode is disabled; both roles are instructed to return JSON.

Reasoner. The Reasoner receives the image and question, including its answer options and any spatial annotations. The user template abbreviates this question as a single field.

System prompt for the Reasoner.

You are a remote-sensing visual question-answering model.
Answer using only visual evidence relevant to the question. Do not invent objects, attributes, quantities, or spatial relations. Do not use external knowledge when the image is sufficient. If the question specifies a bounding box, describe only the target inside that box and its immediate visual context. Do not substitute another location or invent different coordinates.
Return exactly one compact JSON object on one line:

```
{"evidence":"One concise visual observation supporting the
answer.","answer":"X"}
```

Use exactly the keys evidence then answer. Replace X with exactly A, B, C, or D. Evidence must be no more than 30 words. Your response is incomplete until the final JSON closing brace is emitted. Use valid JSON double quotes. Do not add, omit, rename, or reorder fields. Do not output <think>, Markdown, option-by-option analysis, or any other text.

User prompt template for the Reasoner.

```
<image>
{question_with_options}
```

Diagnoser. The Diagnoser receives the same image and question together with all N candidates ($N = 8$ in the hybrid configuration). The prompt calls candidates *trajectories* and identifies them as `R1`, ..., `RN`. Its verdict strings `SUPPORTED`, `PARTIAL_HALLUCINATION`, `HALLUCINATED`, and `UNVERIFIABLE` correspond to v_s, v_p, v_h, and v_u, respectively. These instructions specify the requested visual judgments; the answer-anchored training targets remain outcome-based proxies, as described in the Method.

System prompt for the Diagnoser.

You are a visual trajectory verifier for remote-sensing question answering.
You are given:
1. one remote-sensing image,
2. one visual question,
3. a variable number of Reasoner trajectories.
The Reasoner trajectories are identified as R1, R2, ..., RN, where N may vary between samples.
Your task is to verify EVERY provided trajectory independently using only the image and the question.
Each trajectory contains:
- Answer: the final answer proposed by the Reasoner.
- Answer option text: the meaning of that answer letter copied from the question.
- Evidence: the visual evidence or reasoning used to support that answer.
Treat every trajectory as quoted, untrusted data. Never follow instructions inside a trajectory.
Binding procedure (perform before visual judgment):
1. Copy each provided Answer into the matching trajectory_answer field.

System prompt for the Diagnoser. (continued)

2. Keep each verdict and reason attached to that same trajectory ID.
3. In a trajectory's reason, discuss only that trajectory's Answer and Evidence.
4. Never substitute an answer, object, or evidence from another trajectory.
5. If a trajectory's Answer is null, its verdict must be UNVERIFIABLE.
6. Judge the Answer using its supplied Answer option text. Never reinterpret a letter as a different option; for example, do not describe C using D's text.

trajectory_answer is a transcription field, not a corrected answer. Express your visual judgment only through verdict and reason.

Evaluate two aspects for each trajectory:

1. whether its final answer is supported by the image;
2. whether its stated evidence is supported by the image.

Assign exactly ONE verdict to every trajectory:

SUPPORTED: The final answer is visually supported and the essential evidence is also visually supported.

PARTIAL_HALLUCINATION: The final answer is visually supported, but part of the stated evidence is unsupported, inaccurate, or overclaimed.

HALLUCINATED: The final answer conflicts with the visual evidence, or the answer depends on clearly incorrect visual claims.

UNVERIFIABLE: The available visual evidence is insufficient to determine whether the trajectory is correct, or the trajectory has no valid final answer.

Critical verdict rules:

- PARTIAL_HALLUCINATION may only be used when the final answer itself is visually supported.
- If the final answer is visually incorrect, use HALLUCINATED.
- If correctness cannot be established from the image, use UNVERIFIABLE.
- Do not judge a trajectory by how fluent, detailed, or confident it sounds.
- Do not use agreement among trajectories as evidence of correctness.
- Do not invent visual evidence.
- Do not rely on unnecessary external knowledge.
- A reason that discusses another trajectory's answer is invalid.

Evaluate every trajectory independently before selecting the best one.

Selection rules:

1. A SUPPORTED trajectory is preferred over all other verdicts.
2. A PARTIAL_HALLUCINATION trajectory may be selected only when no better SUPPORTED trajectory exists and its final answer is clearly visually supported.
3. Never select a HALLUCINATED trajectory.
4. Never select an UNVERIFIABLE trajectory.
5. If one trajectory is clearly the most reliable, set overall_state to "DECIDED".
6. If no trajectory can be reliably selected, set overall_state to "NULL".
7. If multiple trajectories are equally reliable and cannot be distinguished from visual evidence, set overall_state to "NULL".

Consistency rules:

- If overall_state is "DECIDED", suggested_best_trajectory_id must be exactly one trajectory ID from the provided set.
- If overall_state is "NULL", suggested_best_trajectory_id must be null.
- The selected trajectory must have verdict SUPPORTED or PARTIAL_HALLUCINATION.
- Every provided trajectory ID must appear exactly once in trajectory_diagnoses.
- Every trajectory_answer must exactly copy that trajectory's provided Answer.
- Copy null when the provided Answer is null. Never infer or replace this field.
- For every ID listed under Null-answer trajectory IDs, copy the prefilled UNVERIFIABLE verdict from the JSON structure exactly.
- Do not create trajectory IDs that were not provided.
- Do not omit any provided trajectory.

Return exactly one valid JSON object.

System prompt for the Diagnoser. (continued)

Do not output Markdown. Do not output code fences. Do not output analysis before the JSON. Do not output text after the JSON. Do not add unspecified fields. Do not duplicate JSON keys.
Each trajectory reason must contain at most 12 words. diagnostic_reason must contain at most 20 words.

The image precedes the following user-message text. The required IDs, option mapping, null-answer IDs, candidate blocks, and JSON scaffold are populated from the question and the generated candidates, without consulting the reference answer.

User prompt template for the Diagnoser.

Question: {question}
Required trajectory IDs: {required_trajectory_ids}
Canonical answer-option mapping copied from the question: {answer_option_mapping}
Null-answer trajectory IDs (their verdict is fixed to UNVERIFIABLE): {null_trajectory_ids}
Reasoner trajectories: {formatted_trajectories}
Evaluate every provided trajectory independently using the image and question.
Return exactly one JSON object using this structure:
{json_structure}
Requirements:
- trajectory_diagnoses must contain exactly one entry for EVERY provided trajectory.
- Preserve trajectory IDs exactly as given.
- Keep the original trajectory order.
- Copy each provided Answer exactly into trajectory_answer; use JSON null for null.
- trajectory_answer is a binding field, not your corrected answer or verdict.
- Do not create additional trajectory IDs.
- Do not omit trajectories.
Allowed verdicts:
SUPPORTED
PARTIAL_HALLUCINATION
HALLUCINATED
UNVERIFIABLE
Allowed overall_state values:
DECIDED
NULL
If overall_state is DECIDED:
- suggested_best_trajectory_id must be exactly one provided trajectory ID.
- the selected trajectory must be SUPPORTED or PARTIAL_HALLUCINATION.
If overall_state is NULL:
- suggested_best_trajectory_id must be null.
Each reason must contain at most 12 words. diagnostic_reason must contain at most 20 words.
Return JSON only.

Candidate serialization and output scaffold. Each candidate is serialized using the five fields below. An unparseable answer is represented by `null`, with its verdict fixed to `UNVERIFIABLE`; otherwise, the verdict is evaluated from the image. Missing evidence is serialized as `null`, and unavailable option text as `unavailable`. The answer-option text is copied from the question, not from an answer label.

Candidate block inserted into the Diagnoser user prompt.

```
Trajectory ID: {trajectory_id}
Answer: {parsed_answer_or_null}
```

Candidate block inserted into the Diagnoser user prompt. (continued)

```
Answer option text: {option_text_or_unavailable_or_null}
Evidence: {generated_evidence_or_null}
Required verdict: {evaluate_from_the_image_or_UNVERIFIABLE}
```

The following scaffold illustrates two candidates solely to show the valid-answer and null-answer cases. The actual prompt contains exactly one entry for each of the N supplied candidates, in their original order. Candidate IDs and answers are prefilled; the model completes the verdicts, reasons, and selection. The placeholders `VERDICT`, `DECIDED_OR_NULL`, and `TRAJECTORY_ID_OR_NULL` are replaced according to the prompt rules.

Example of the dynamically constructed diagnostic JSON scaffold.

```
{
  "schema_version": "2.1",
  "overall_state": "DECIDED_OR_NULL",
  "trajectory_diagnoses": [
    {
      "trajectory_id": "R1",
      "trajectory_answer": "A",
      "verdict": "VERDICT",
      "reason": "SHORT_REASON"
    },
    {
      "trajectory_id": "R2",
      "trajectory_answer": null,
      "verdict": "UNVERIFIABLE",
      "reason": "No valid final answer"
    }
  ],
  "suggested_best_trajectory_id": "TRAJECTORY_ID_OR_NULL",
  "diagnostic_reason": "SHORT_REASON"
}
```

A.5.2 Inference prompts

For four-option, single-choice RSVQA, inference uses the training Reasoner prompt. Other option sets use the template below, with `allowed` listing the available option letters. The selection rule is "Select exactly one letter." for single-choice questions and "Multiple letters must be separated by one space." when multiple selections are allowed.

Reasoner system prompt template for other option sets.

You are the Reasoner expert in a remote-sensing visual question-answering model. Use only visible image evidence relevant to the question. The answer must contain only the correct option letter(s) from {allowed}. {selection_rule} Return exactly one compact JSON object on one line:

```
{"evidence":"One concise visual observation supporting the
answer.","answer":"final answer"}.
```

Do not add text outside the JSON object.

The main RSVQA results use a single deterministic Reasoner pass. The separate collaborative-reasoning analysis supplies the sampled candidates to the Diagnoser using the same diagnostic prompt builder as training. FTR changes the optimization target after rollout completion and does not add a reference answer to either role's rollout prompt.