RS-OPSD: RELIABLE PRIVILEGED ON-POLICY SELF-DISTILLATION FOR ULTRA-HIGH-RESOLUTION REMOTE SENSING VQA

Anonymous authors
Paper under double-blind review

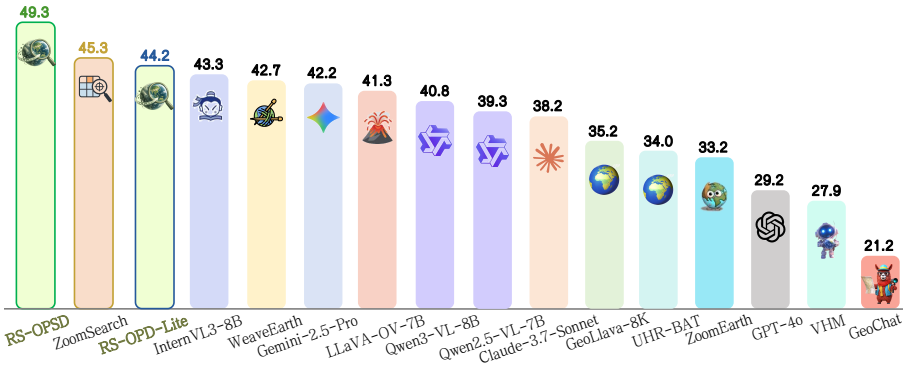


Figure 1: Average scores on Ultra-High-Resolution remote sensing VQA benchmarks, including XLRS-Bench, MME-RealWorld-RS, and LRS-VQA.

ABSTRACT

Ultra-high-resolution (UHR) remote sensing visual question answering (VQA) requires models to resolve small visual evidence within extremely large images. Existing approaches typically rely on token pruning, visual search, or tool-augmented reasoning at inference time. We instead investigate whether the benefit of *zoom-in visual privilege* can be internalized into the model. We introduce RS-OPSD, a reliable privileged on-policy self-distillation (OPSD) framework for UHR remote sensing VQA. To provide high-quality privileged information with explicit question-relevant evidence, we construct **GeoEvidence-6K**, containing 6,750 VQA samples across seven task categories with evidence-region annotations, and develop **Human Feedback-Guided Skill Refinement (HF-SR)** for scalable annotation. To address context loss from tight crops and conflicting signals from imperfect teachers, RS-OPSD introduces **Context-Preserving Visual Privilege (CPVP)** and **Correctness-Aligned Distillation (CAD)**. Without any additional visual search or tool calls at inference time, RS-OPSD achieves state-of-the-art (SOTA) performance on XLRS-Bench, MME-RealWorld-RS, and LRS-VQA, outperforming pervious SOTA models of comparable scale by an average of 4.0 percentage points. Moreover, our 2B variant, RS-OPD-*Lite*, surpasses most 8B-scale models while achieving the fastest measured inference speed.

1 INTRODUCTION

UHR remote sensing VQA requires models to reason over images containing tens of millions of pixels, where the evidence is often confined to only a tiny fraction of the scene. Although modern vision-language foundation models (Bai et al., 2025b;a; Li et al., 2024; Zhu et al., 2025) exhibit strong semantic understanding, their effective perceptual capacity is easily overwhelmed by such extremely large visual inputs. Recent studies such as ZoomSearch (Zhou et al., 2025) and Vision-OPD (Yuan et al., 2026) suggest that many fine-grained failures are not simply caused by an inability

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.

to recognize the target itself, but by a more fundamental *where-to-look* bottleneck—the model often fails to effectively attend to and exploit question-relevant visual evidence in extremely large images.

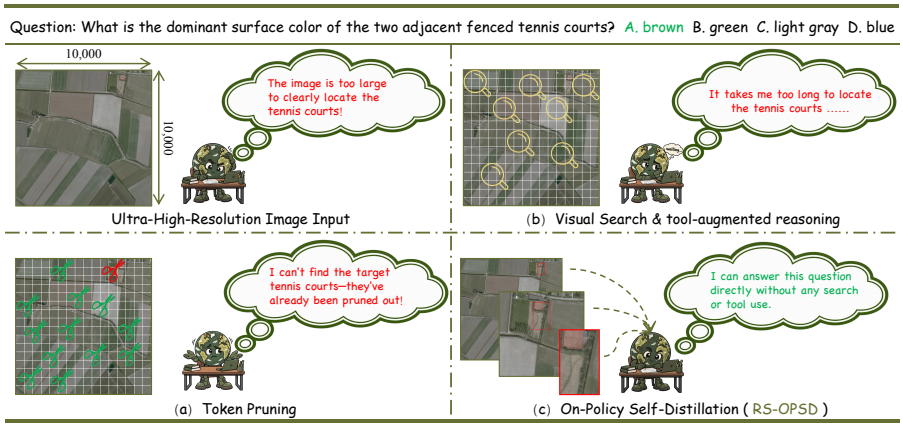


Figure 2: Comparison of representative paradigms for UHR remote sensing VQA.

As illustrated in Figure 2, existing approaches mainly address this bottleneck from three directions. First, *token-pruning* methods exploit the high redundancy of UHR imagery by retaining only a subset of visual tokens (Luo et al., 2025; Dang et al., 2026). However, hard pruning is typically difficult to reverse: once evidence token is discarded, it is no longer available to subsequent reasoning. Second, *visual search* methods search over image patches to identify relevant regions before answering (Zhou et al., 2025; Ma et al., 2026). While effective, such search introduces additional inference-time computation. Third, *tool-augmented reasoning* methods allow the model to actively invoke zoom-in tools during inference, often through multiple rounds of refinement (Liu et al., 2026; Wang et al., 2026). These methods demonstrate the benefit of exposing the model to targeted high-resolution views, but require additional tool calls and visual search. This observation motivates a different question: *Can the privilege brought by visual zooming be internalized into the model during training, such that it can exploit the same fine-grained visual prior without additional search or tool use?*

A natural way to realize this idea is *on-policy self-distillation* (OPSD) (Zhao et al., 2026). Instead of introducing additional visual operations at inference time, OPSD equips a teacher with privileged visual information during training and transfers this advantage to a student operating under the standard input. In UHR remote sensing VQA, zoomed-in evidence can naturally serve as such visual privilege: during training, the teacher is exposed to fine-grained question-relevant views, while the student learns to better exploit the corresponding visual evidence from the standard full-image input.

Realizing this training paradigm requires high-quality zoom-in evidence as visual privilege. We therefore construct **GeoEvidence-6K**, a UHR remote sensing VQA dataset containing 6,750 samples across seven task categories, with explicit annotations for question-relevant visual evidence. Since accurately annotating such regions requires careful inspection of extremely large images, we further develop **Human Feedback-Guided Skill Refinement (HF-SR)**, which progressively incorporates human feedback into reusable annotation skills to improve subsequent data construction.

We then conduct a pilot study using full-image inputs for the student and evidence-crop privileges for a teacher initialized from the same model. However, this straightforward formulation yields only marginal gains. We identify two major bottlenecks: first, the crop-only privilege provides only a modest advantage over the student; second, the privileged teacher itself remains imperfect, which can introduce substantial conflicting or even harmful supervision. To address these issues, we introduce two complementary designs in RS-OPSD. **Context-Preserving Visual Privilege (CPVP)** equips the teacher with global, contextual, and fine-grained evidence views, strengthening the privileged signal while preserving surrounding context. **Correctness-Aligned Distillation (CAD)** filters teacher supervision through sample-level reliability gating and retains only token-level update directions aligned with the correctness of the student response.

Using Qwen3-VL-8B (Bai et al., 2025a) as the base model, RS-OPSD achieves SOTA performance on all three UHR remote sensing VQA benchmarks, outperforming the strongest competing methods

by 1.6 percentage points on XLRS-Bench (Wang et al., 2025b), 3.9 points on MME-RealWorld-RS (Zhang et al., 2025), and 2.0 points on LRS-VQA (Luo et al., 2025). We further instantiate the same framework with a Qwen3-VL-2B student and a privileged Qwen3-VL-8B teacher, yielding RS-OPD-*Lite*. Despite retaining only the 2B student at inference time, RS-OPD-*Lite* surpasses most 8B-scale models. Moreover, since neither variant requires additional visual search or tool calls, they both achieve low inference latency, with RS-OPD-*Lite* achieving the lowest inference latency, 12.8% lower than the second-fastest method. Our main contributions are summarized as follows:

- We introduce RS-OPSD, which formulates UHR remote sensing VQA as an on-policy self-distillation problem and internalizes the priviledge of zoom-in evidence into the model.
- To support this paradigm, we construct **GeoEvidence-6K**, a high-quality UHR remote sensing VQA dataset with explicit evidence-region annotations. We further propose **Human Feedback-Guided Skill Refinement**, which iteratively converts accumulated human feedback into reusable annotation skills for subsequent data construction.
- We develop two complementary components for reliable privileged distillation: **Context-Preserving Visual Privilege**, which provides the teacher with global, contextual, and fine-grained visual evidence, and **Correctness-Aligned Distillation**, which suppresses unreliable teacher supervision through sample-level reliability gating and correctness-aligned token updates.
- Extensive experiments show that RS-OPSD achieves SOTA performance while maintaining efficient, search-free and tool-free inference. Moreover, RS-OPD-*Lite* retains competitive accuracy and achieves the fastest inference speed among all compared methods.

2 Related Work

As shown in Figure 2, existing approaches to UHR remote sensing VQA mainly follow three paradigms: *token pruning*, *visual search*, and *tool-augmented reasoning*.

Token-pruning methods exploit the redundancy in UHR remote sensing imagery by retaining question-relevant visual tokens while discarding less informative regions. GeoLLaVA-8K (Wang et al., 2025a) combines Background Token Pruning with Anchored Token Selection to suppress redundant background tokens while preserving object-centric visual tokens. UHR-BAT (Dang et al., 2026) further introduces query-guided multi-scale token selection and region-faithful preserve-and-merge strategies to allocate a fixed token budget toward question-relevant regions.

Visual search methods explicitly identify informative regions and construct compact evidence from the original UHR image before reasoning, avoiding dense processing of the entire scene. Zoom-Search (Zhou et al., 2025) performs hierarchical zoom search to locate question-relevant patches and reassembles the selected evidence with spatial layout preserved. WeaveEarth (Ma et al., 2026) constructs a compact Minimal Support Evidence Set through global-aware evidence selection and integrates local evidence, spatial metadata, and relative topology for subsequent reasoning.

Tool-augmented reasoning methods equip VLMs with explicit visual tools, enabling active acquisition of high-resolution evidence during inference. ZoomEarth (Liu et al., 2026) introduces an active perception paradigm that learns adaptive cropping and zooming through supervised fine-tuning and reinforcement learning, allowing the model to revisit informative regions during reasoning. Building on this direction, GeoEyes (Wang et al., 2026) further addresses the tendency of zoom-enabled models to follow homogeneous tool-use patterns, learning more adaptive on-demand zooming and stopping behaviors through staged supervised and reinforcement learning.

3 GeoEvidence-6K

To provide the teacher with zoom-in visual privileges, we require high-quality UHR remote sensing data pairing each question with a target region. We therefore construct **GeoEvidence-6K**, a manually verified UHR remote sensing VQA dataset with explicit target-region annotations. Since annotating these regions is labor-intensive and prone to repetitive errors, we introduce **Human Feedback-Guided Skill Refinement (HF-SR)**, a human-in-the-loop annotation framework that consolidates human corrections into reusable annotation skills. Human feedback thus not only corrects individual samples but also continuously improves subsequent annotation.

3.1 Dataset Overview

GeoEvidence-6K is constructed from 1,050 UHR remote sensing images collected from five public data sources, as summarized in Table 1. The images average $10{,}714 \times 10{,}714$ pixels, with ground sampling distances (GSDs) ranging from 0.08 m to 0.50 m, and span diverse urban, agricultural, natural, coastal, and mountainous scenes. The annotated target regions average only $1{,}095 \times 1{,}074$ pixels, roughly 1% of the full-image area. This large scale disparity highlights the difficulty of locating and resolving question-relevant visual evidence in UHR scenes.

Table 1: Statistics of the image sources used in GeoEvidence-6K.

Source	Nums.	Image Size	Target Size	GSD	Country
MiniFrance (Castillo-Navarro et al., 2022)	110	$10K$	509×517	0.50 m	France
HRSCD (Daudt et al., 2019)	40	$10K$	516×517	0.50 m	France
SWISSIMAGE (Federal Office of Topography swisstopo, 2024)	300	$10K$	837×817	0.10 m	Switzerland
GeoNRW (Geobasis NRW, 2025)	300	$10K$	997×987	0.10 m	Germany
Beeldmateriaal (Beeldmateriaal Nederland, 2025)	300	$12K$	$1{,}741 \times 1{,}696$	0.08 m	Netherlands

The dataset covers seven multiple-choice task categories. Five are local-evidence tasks, including `counting`, `position`, `color`, `category`, and `shape`, while `land-use` and `route-planning` are global-context tasks. For each source image, we select diverse target regions and annotate questions of different task types, resulting in a total of 6,750 training samples. Representative annotations are shown in Figure 3.

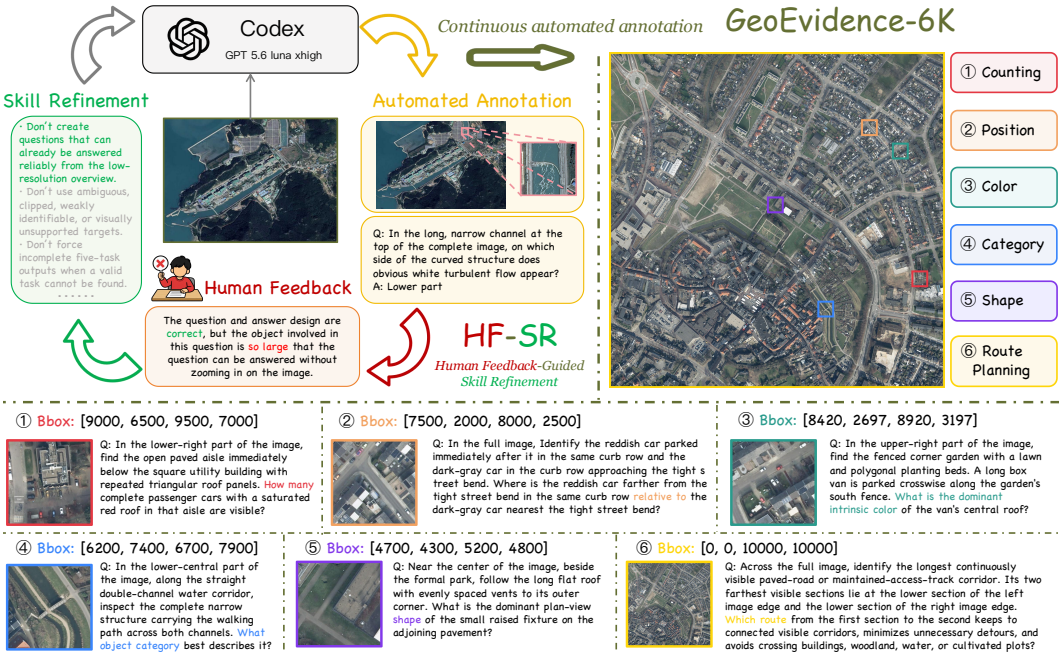


Figure 3: Overview of the HF-SR pipeline, together with representative cases of GeoEvidence-6K.

3.2 HF-SR: Human Feedback-Guided Skill Refinement

Figure 3 illustrates our HF-SR pipeline. Let $\mathcal{A}$ denote a fixed annotation agent and $\mathcal{S}_k$ the annotation skill at refinement stage k. Given a UHR remote sensing image x_i, the agent generates $\hat{y}_i = \mathcal{A}(x_i; \mathcal{S}_k)$, which is reviewed by humans to obtain feedback $f_i = \mathcal{H}(x_i, \hat{y}_i)$. Unlike conventional human verification, HF-SR reuses accumulated feedback to update the annotation skill. Specifically, feedback collected between two checkpoints is $\mathcal{F}_k = \{f_i \mid c_k < i \leq c_{k+1}\}$, and the skill is refined as $\mathcal{S}_{k+1} = \mathcal{R}(\mathcal{S}_k, \mathcal{F}_k)$, where $\mathcal{R}$ denotes an agent-driven refinement process that summarizes recurring errors and human corrections into reusable procedural rules.

The agent follows an overview-to-detail procedure: it first inspects a downsampled full-image overview and then examines candidate regions at native resolution. For local-evidence tasks, it selects spatially diverse target regions and generates the corresponding bounding boxes, questions, and reference answers. Human reviewers jointly inspect the full image, target regions, and structured annotations, correcting issues such as ambiguous references, incomplete target regions, invalid crop geometry, or inconsistent question–answer design.

To account for the evolving maturity of the annotation skill, we schedule refinement checkpoints using cumulative Fibonacci intervals, with $c_k = \sum_{j=1}^{k} F_j$ and $F_1 = F_2 = 1$. Frequent updates are used during the cold-start stage, when new failure patterns emerge rapidly, while progressively longer intervals are adopted as the skill stabilizes to amortize refinement cost.

4 PRELIMINARY

On-Policy Distillation. On-policy distillation (OPD) transfers knowledge from a teacher policy π_T to a student policy π_θ on trajectories sampled from the student itself. Given an input x, the student first generates $\mathbf{y} = (y_1, \ldots, y_T) \sim \pi_\theta(\cdot \mid x)$. At each student-visited prefix $\mathbf{y}_{<t}$, the student and teacher produce next-token distributions:

$$p_t(\cdot) = \pi_\theta(\cdot \mid x, \mathbf{y}_{<t}), \qquad q_t(\cdot) = \pi_T(\cdot \mid x, \mathbf{y}_{<t}). \tag{1}$$

OPD then minimizes the reverse KL divergence along the student trajectory:

$$\mathcal{L}_{\mathrm{OPD}} = \mathbb{E}_{\mathbf{y}\sim\pi_\theta(\cdot|x)} \left[\sum_{t=1}^{T} D_{\mathrm{KL}} \left(p_t \,\|\, q_t \right) \right]. \tag{2}$$

Unlike off-policy distillation, the teacher provides supervision on states actually visited by the current student policy, reducing train–inference distribution mismatch.

On-Policy Self-Distillation. On-policy self-distillation (OPSD) removes the need for a separate stronger teacher by assigning different contextual roles to the same underlying model (Zhao et al., 2026). The student observes the standard input x, whereas the teacher is additionally conditioned on privileged information z that is available only during training:

$$p_t^S = \pi_\theta(\cdot \mid x, \mathbf{y}_{<t}), \qquad p_t^T = \pi_{\theta_T}(\cdot \mid x, z, \mathbf{y}_{<t}). \tag{3}$$

The privileged teacher evaluates the same student-generated trajectory, and the student is optimized toward its distribution:

$$\mathcal{L}_{\mathrm{OPSD}} = \mathbb{E}_{\mathbf{y}\sim\pi_\theta(\cdot|x)} \left[\sum_{t=1}^{T} D_{\mathrm{KL}} \left(p_t^S \,\|\, \mathrm{sg}\big[p_t^T\big] \right) \right], \tag{4}$$

where $\mathrm{sg}[\cdot]$ denotes stop-gradient. In this way, OPSD transfers capabilities exposed by privileged training information into the student while retaining the standard input at inference time.

5 RS-OPSD

5.1 MOTIVATION

We first conduct a pilot study to examine whether zoom-in evidence can be directly exploited through OPSD for UHR remote sensing VQA. As shown in Table 3, LoRA fine-tuning on GeoEvidence-6K improves the baseline from 40.8 to 45.7 on average, whereas directly applying OPSD with the evidence crop as the teacher privilege achieves only 45.2. Although OPSD still improves over the base model, it fails to outperform standard LoRA fine-tuning, suggesting that crop-based privileged distillation is not yet sufficiently effective in this setting.

To understand this limitation, we further evaluate the teacher–student capability gap on GeoEvidence-6K under direct-answer evaluation. Using the same Qwen3-VL-8B model, the full-image input achieves 63.85% accuracy, while the evidence crop reaches 69.91%. This diagnostic reveals two bottlenecks: first, the crop privilege provides only a 6.06-point advantage, resulting

in a relatively weak privileged signal; second, the privileged model itself remains imperfect, with nearly 30% of the training samples answered incorrectly, which can introduce conflicting supervision. These observations motivate two complementary designs in RS-OPSD: **Context-Preserving Visual Privilege (CPVP)** enriches the teacher with additional contextual and global information to strengthen its privileged advantage, while **Correctness-Aligned Distillation (CAD)** filters unreliable teacher signals and retains only update directions aligned with verified answer correctness.

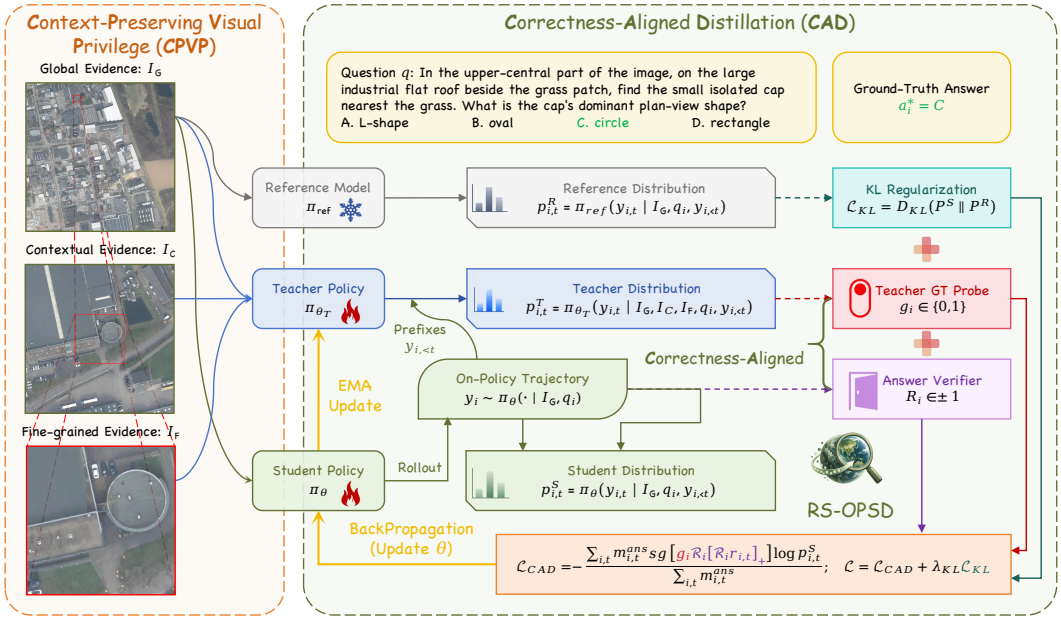


Figure 4: Overview of RS-OPSD, consisting of CPVP and CAD.

5.2 CPVP: What Should the Teacher See?

Building on the OPSD formulation in Section 4, each training sample contains a UHR remote sensing image I_i, a question q_i, a ground-truth answer a_i^*, and an annotated evidence region B_i. In our setting, the standard input x corresponds to the global evidence I_G, while the privileged information z is constructed from the annotated evidence region and provided only to the teacher.

A tight evidence crop improves target visibility but may discard surrounding context. Conversely, the global image preserves scene-level information but weakens fine-grained targets. As shown in Figure 4, we therefore propose **CPVP**, which provides the teacher with three complementary views: Global Evidence I_G, Contextual Evidence I_C, and Fine-grained Evidence I_F. Specifically, I_F is the tight crop corresponding to B_i, while I_C is obtained by keeping the same crop center and doubling its width and height to include surrounding context. The resulting privileged information is $z_i = \{I_{C,i}, I_{F,i}\}$. The student observes the standard input $(I_{G,i}, q_i)$, while the teacher is additionally conditioned on the privileged information $z_i = \{I_{C,i}, I_{F,i}\}$. CPVP thus strengthens the teacher privilege with fine-grained evidence and surrounding context while preserving the global input.

5.3 CAD: When Should the Student Trust It?

Although CPVP provides the teacher with richer visual evidence, privileged information does not guarantee reliable supervision. We therefore propose **CAD**, which filters privileged supervision at both the sample and token levels. Following the on-policy formulation in Section 4, the privileged teacher evaluates the same student-generated trajectory. For each sampled token $y_{i,t}$, we define the teacher–student preference as

$$r_{i,t} = \mathrm{sg}\left[\log p_{i,t}^T - \log p_{i,t}^S\right], \tag{5}$$

where $p_{i,t}^S$ and $p_{i,t}^T$ denote the probabilities assigned to the sampled token $y_{i,t}$ under the student and privileged-teacher distributions defined in Section 4, respectively.

Teacher reliability gate. Before using these token-level preferences, CAD first evaluates whether the privileged teacher provides reliable supervision for the current sample. Given the GT response a_i^*, we perform a teacher-forced prefix probe and define

$$g_i = \prod_k \mathbf{1}\left[\tau_{i,k}^* = \arg\max_{u \in \mathcal{C}_{i,k}} \pi_{\theta_T}\left(u \mid I_{G,i}, z_i, q_i, \boldsymbol{\tau}_{i,<k}^*\right)\right], \tag{6}$$

where $\boldsymbol{\tau}_i^*$ denotes the tokenized GT response and $\mathcal{C}_{i,k}$ denotes the valid continuations at position k. Thus, $g_i = 1$ only when the teacher remains aligned with the GT response throughout the probe.

Correctness-aligned token filtering. We further use the correctness of the student response to determine which teacher preferences should be retained:

$$R_i = \begin{cases} +1, & \hat{a}_i = a_i^*, \\ -1, & \hat{a}_i \neq a_i^*. \end{cases} \tag{7}$$

The resulting CAD signal is

$$A_{i,t} = g_i R_i \left[R_i r_{i,t}\right]_+ . \tag{8}$$

When the student response is correct, CAD retains teacher preferences that reinforce the sampled tokens; when the response is incorrect, it retains preferences that suppress them. Teacher signals that are unreliable or inconsistent with verified answer correctness are discarded.

The CAD objective is

$$\mathcal{L}_{\mathrm{CAD}} = -\frac{\sum_{i,t} m_{i,t}^{\mathrm{ans}} \, \mathrm{sg}[A_{i,t}] \log p_{i,t}^S}{\sum_{i,t} m_{i,t}^{\mathrm{ans}}}, \tag{9}$$

where $m_{i,t}^{\mathrm{ans}}$ is a binary mask selecting valid answer tokens.

Finally, we regularize the student toward a frozen reference model that receives the same standard input $(I_{G,i}, q_i)$ as the student. The overall training objective is

$$\mathcal{L} = \mathcal{L}_{\mathrm{CAD}} + \lambda_{\mathrm{KL}} \mathcal{L}_{\mathrm{KL}}, \tag{10}$$

where λ_{KL} controls the strength of reference regularization. We further apply gradient-norm clipping with threshold c_{grad} to stabilize optimization.

6 Experiments

6.1 Experimental Settings

Model Training. RS-OPSD is initialized from Qwen3-VL-8B-Instruct and trained exclusively on GeoEvidence-6K. We set the KL regularization coefficient $\lambda_{\mathrm{KL}} = 1 \times 10^{-3}$, the gradient clipping threshold $c_{\mathrm{grad}} = 5$, and the global batch size to 96. Training lasts for 150 steps. The teacher is initialized from the student and updated via exponential moving average (EMA) throughout training. For RS-OPD-*Lite*, the student is initialized from Qwen3-VL-2B-Instruct, while Qwen3-VL-8B-Instruct serves as the privileged teacher and remains frozen during training. All other settings follow RS-OPSD, except that RS-OPD-*Lite* is trained for 120 steps.

Benchmarks. We evaluate our models on three UHR remote sensing VQA benchmarks: XLRS-Bench (Wang et al., 2025b), MME-RealWorld-RS (Zhang et al., 2025), and LRS-VQA (Luo et al., 2025). XLRS-Bench has an average image resolution of $8{,}500 \times 8{,}500$ pixels and evaluates fine-grained perception and reasoning tasks. MME-RealWorld-RS has an average resolution of $5{,}602 \times 4{,}445$ pixels and contains 3,738 single-choice questions covering color, counting, and position. LRS-VQA contains 1,657 images with an average resolution of $7{,}099 \times 6{,}329$ pixels and 7,333 question–answer pairs across eight categories. We follow the official evaluation protocols of each benchmark and report the corresponding task-averaged scores.

Baselines. We compare against 14 representative baselines spanning four categories. (1) *Closed-source VLMs*: GPT-4o, Claude 3.7 Sonnet, and Gemini 2.5 Pro. (2) *Open-source VLMs*: LLaVA-OV-7B, Qwen2.5-VL-7B, Qwen3-VL-8B, and InternVL3-8B. (3) *Remote sensing VLMs*: GeoChat and VHM. (4) *Specialized methods for UHR remote sensing VQA*, including token-pruning approaches such as GeoLLaVA-8K and UHR-BAT, visual-search approaches such as ZoomSearch and WeaveEarth, and tool-augmented reasoning approaches represented by ZoomEarth.

Table 2: Quantitative comparison results. Green, yellow, and blue denote the first-, second-, and third-best results. Speed is the average inference time across the three benchmarks (s/sample).

Method	Pub.	Backbone	Param.	XLRS	MME	LRS	Avg.	Speed
Closed-Source Vision Language Models								
GPT-4o (Hurst et al., 2024)	–	–	–	32.4	28.2	27.1	29.2	-
Claude 3.7 Sonnet (Cla)	–	–	–	40.5	45.7	28.4	38.2	-
Gemini 2.5 Pro (Comanici et al., 2025)	–	–	–	44.1	51.7	**30.9**	42.2	-
Open-Source Vision Language Models								
LLaVA-OV-7B (Li et al., 2024)	TMLR'25	LLaVA-OV	7B	43.4	53.3	27.3	41.3	2.38
Qwen2.5-VL-7B (Bai et al., 2025b)	arXiv'25	Qwen2.5-VL	7B	47.4	41.0	29.4	39.3	1.87
Qwen3-VL-8B (Bai et al., 2025a)	arXiv'25	Qwen3-VL	8B	**50.5**	41.9	30.1	40.8	1.75
InternVL3-8B (Zhu et al., 2025)	arXiv'25	InternVL3	8B	45.6	55.0	29.2	43.3	**1.25**
Remote Sensing Vision Language Models								
GeoChat (Kuckreja et al., 2024)	CVPR'24	LLaVA-v1.5	7B	22.9	21.3	19.5	21.2	1.33
VHM (Pang et al., 2025)	AAAI'25	Vicuna	7B	32.8	24.1	26.7	27.9	2.64
Specialized Methods for UHR Remote Sensing VQA								
GeoLLaVA-8K (Wang et al., 2025a)	NIPS'25	LongVA	7B	**51.5**	28.4	25.8	35.2	**1.30**
UHR-BAT (Dang et al., 2026)	ICML'26	LongVA	7B	48.6	33.3	20.0	34.0	7.40
ZoomSearch (Zhou et al., 2025)	arXiv'25	LLaVA-OV	7B	48.0	**57.6**	30.2	**45.3**	43.5
WeaveEarth (Ma et al., 2026)	MM'26	Qwen3-VL	8B	49.5	47.2	**31.3**	42.7	2.94
ZoomEarth (Liu et al., 2026)	CVPR'26	Qwen2.5-VL	3B	39.2	38.7	21.6	33.2	14.2
Ours								
RS-OPD-*Lite*	–	Qwen3-VL	2B	45.8	**56.2**	30.5	**44.2**	**1.09**
RS-OPSD	–	Qwen3-VL	8B	**53.1**	**61.5**	**33.3**	**49.3**	1.58

6.2 Comparison with State-of-the-Art Methods

Overall performance. As shown in Table 2, RS-OPSD achieves the best performance on all three benchmarks, reaching 53.1 on XLRS-Bench, 61.5 on MME-RealWorld-RS, and 33.3 on LRS-VQA, with an average score of 49.3, exceeding the strongest competing method by 4.0 points. Detailed category-wise results and analyses are provided in Appendix D.

Accuracy–efficiency trade-off. RS-OPSD runs at 1.58 s/sample, even faster than Qwen3-VL-8B baseline, while avoiding the substantial overhead of visual search and tool calls. More notably, RS-OPD-*Lite* achieves an average score of 44.2 with only 2B parameters, outperforming most 8B-scale models. It is also the fastest method at 1.09 s/sample, with 12.8% lower latency than the second-fastest method. These results demonstrate that internalizing privileged visual information through OPSD can improve UHR remote sensing VQA while preserving both accuracy and efficiency.

6.3 Ablation Studies and Analysis

Core design analysis. Table 3 evaluates the contribution of the core designs in RS-OPSD. LoRA (Hu et al., 2021) fine-tuning on GeoEvidence-6K improves the average score of Qwen3-VL-8B from 40.8 to 45.7, indicating the high quality of the **GeoEvidence-6K**. In contrast, directly applying OPSD with crop-only privilege in our pilot setting slightly reduces the average score to 45.2, confirming that naïve privileged distillation is limited by the insufficient teacher–student gap and imperfect teacher supervision. Introducing CPVP raises the average score to 47.2, while CAD independently improves it to 47.0, showing that richer visual privilege and correctness-aligned supervision both contribute positively to mitigating these two bottlenecks. Combining CPVP and CAD further yields the best performance of 49.3, reaching the state-of-the-art results reported in Table 2.

Hyperparameter analysis. Table 4 examines the effects of the KL coefficient λ_{KL} and gradient clipping threshold c_{grad}. KL regularization consistently improves over the setting with $\lambda_{\mathrm{KL}} = 0$, with $\lambda_{\mathrm{KL}} = 10^{-3}$ achieving the best average score of 49.0. Increasing λ_{KL} to 10^{-2} slightly degrades performance, suggesting that overly strong regularization may constrain adaptation to the privileged

Table 3: Ablation study of the core designs in RS-OPSD.

Core Design			Label	Benchmarks			
OPSD	CPVP	CAD		XLRS	MME-RW-RS	LRS-VQA	Avg.
Qwen3-VL-8B-Instruct			Base Model	50.5	41.9	30.1	40.8
✗	✗	✗	LoRA Finetuning on GeoEvidence-6K	49.9	55.2	**32.1**	45.7
✓	✗	✗	Naïve OPSD in Subsec. 5.1	50.5	53.8	31.3	45.2
✓	✓	✗	OPSD + CPVP	**51.3**	**59.0**	31.2	**47.2**
✓	✗	✓	OPSD + CAD	**51.2**	**58.3**	**31.5**	**47.0**
✓	✓	✓	RS-OPSD	**53.1**	**61.5**	**33.3**	**49.3**

Table 4: Hyperparameter studies of KL regularization and gradient clipping.

KL Coefficient		Benchmarks				Clipping Threshold		Benchmarks			
c_{grad}	λ_{KL}	XLRS	MME-RW-RS	LRS-VQA	Avg.	c_{grad}	λ_{KL}	XLRS	MME-RW-RS	LRS-VQA	Avg.
1	0	51.3	59.0	31.2	47.2	5	1×10^{-3}	**53.1**	**61.5**	**33.3**	**49.3**
1	1×10^{-2}	51.2	**62.0**	32.8	**48.7**	20	1×10^{-3}	51.0	60.0	**34.0**	48.3
1	1×10^{-3}	**53.2**	**60.8**	33.1	**49.0**	100	1×10^{-3}	**51.5**	59.9	**33.4**	48.3

teacher. With $\lambda_{\text{KL}} = 10^{-3}$ fixed, $c_{\text{grad}} = 5$ performs best at 49.3, while $c_{\text{grad}} = 20$ and 100 both decrease the average to 48.3. We therefore use $\lambda_{\text{KL}} = 10^{-3}$ and $c_{\text{grad}} = 5$ in all main experiments.

Training dynamics analysis. Figure 5 further examines the training dynamics of RS-OPSD over three epochs. Figure 5(a) tracks the accuracy of the student and EMA teacher. Both improve steadily throughout training, while the teacher consistently remains stronger and the teacher–student gap gradually narrows, indicating progressive transfer of privileged knowledge to the student. Figure 5(b) shows that suppressive token activation steadily decreases, while reinforced and overall active tokens drop sharply in the third epoch, suggesting that fewer teacher signals remain informative as the student approaches the teacher. Finally, Figure 5(c) reports task-wise student accuracy. Nearly all task categories exhibit consistent improvements across training.

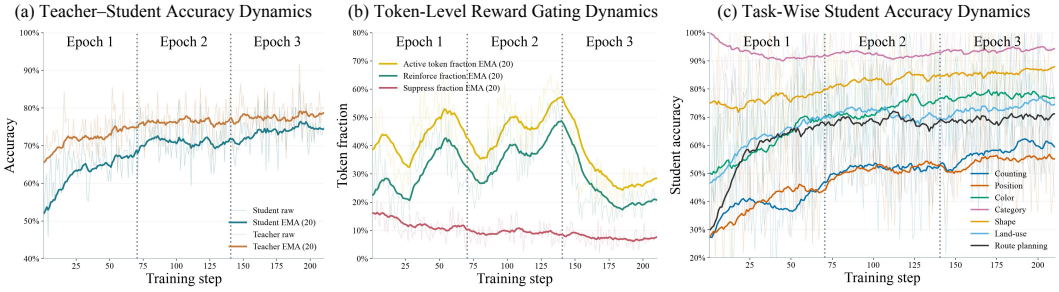


Figure 5: Training dynamics of RS-OPSD across three epochs.

7 Conclusion

In this work, we present RS-OPSD, a reliable privileged OPSD framework for UHR remote sensing VQA. Instead of relying on additional visual search or tool use at inference time, RS-OPSD internalizes zoom-in visual privilege into the model during training. To support this paradigm, we construct **GeoEvidence-6K** with explicit evidence-region annotations and introduce **HF-SR** for human-feedback-guided data construction. We further develop **CPVP** to provide richer contextual and fine-grained visual privilege, and **CAD** to suppress unreliable teacher guidance. Extensive experiments demonstrate that RS-OPSD achieves SOTA performance while maintaining efficient inference, and that RS-OPD-*Lite* further offers a favorable accuracy–efficiency trade-off with a substantially smaller student model. We further discuss the limitations of the current framework and potential directions for future work in Appendix E.

AI USE STATEMENT

In this work, we used generative AI tools as part of the construction of GeoEvidence-6K. Specifically, a fixed Codex agent powered by GPT-5.6 Luna with `xhigh` reasoning effort was used to generate and inspect candidate annotations under the Human Feedback-Guided Skill Refinement (HF-SR) pipeline. All generated annotations were manually reviewed, and human corrections were incorporated into reusable annotation skills for subsequent data construction.

Generative AI tools were also used to assist with language refinement and LaTeX formatting. All experiments were designed and conducted by the human authors, and all scientific conclusions and mathematical formulations were independently developed by the authors.

ETHICS STATEMENT

This work uses publicly available remote sensing imagery and does not involve human-subject experiments or personally identifiable information. GeoEvidence-6K is constructed from publicly released datasets and national mapping programs, with annotations restricted to remote sensing VQA content. Human involvement is limited to annotation verification and correction. We are not aware of additional ethical concerns associated with the proposed method or its intended use.

REPRODUCIBILITY STATEMENT

The main paper and appendix provide detailed descriptions of the training and evaluation procedures, hyperparameter settings, and corresponding ablation studies. An anonymized version of the training and evaluation code is available at `https://anonymous.4open.science/r/RS-OPSD`. We will release GeoEvidence-6K and the model weights for RS-OPSD and RS-OPD-*Lite* upon publication.

REFERENCES

Claude 3.7 sonnet system card. URL `https://api.semanticscholar.org/CorpusID:276612236`.

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

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, 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*, 2025b.

Beeldmateriaal Nederland. High-resolution aerial imagery of the netherlands. `https://www.beeldmateriaal.nl/dataroom`, 2025. Accessed September 5, 2026.

Javiera Castillo-Navarro, Bertrand Le Saux, Alexandre Boulch, Nicolas Audebert, and Sébastien Lefèvre. Semi-supervised semantic segmentation in earth observation: The MiniFrance suite, dataset analysis and multi-task network study. *Machine Learning*, 111(9):3125–3160, 2022. doi: 10.1007/s10994-020-05943-y.

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.

Yunkai Dang, Minxin Dai, Yuekun Yang, Zhangnan Li, Wenbin Li, Feng Miao, and Yang Gao. Uhrbat: Budget-aware token compression vision-language model for ultra-high-resolution remote sensing. In *Forty-third International Conference on Machine Learning*, 2026.

Rodrigo Caye Daudt, Bertrand Le Saux, Alexandre Boulch, and Yann Gousseau. Multitask learning for large-scale semantic change detection. *Computer Vision and Image Understanding*, 187: 102783, 2019. doi: 10.1016/j.cviu.2019.07.003.

Federal Office of Topography swisstopo. SWISSIMAGE 10 cm: The digital color orthophotomosaic of switzerland. https://www.swisstopo.admin.ch/en/orthoimage-swissimage-10, 2024. Accessed September 5, 2026.

Geobasis NRW. Digital orthophotos of north rhine-westphalia (DOP). https://registry.gdi-de.org/id/de.nw/DOP, 2025. Accessed September 5, 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. *arXiv preprint arXiv:2106.09685*, 2021.

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 *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 27831–27840, 2024.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles*, 2023.

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

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 (CVPR)*, 2026.

Junwei Luo, Yingying Zhang, Xue Yang, Kang Wu, Qi Zhu, Lei Liang, Jingdong Chen, and Yansheng Li. When large vision-language model meets large remote sensing imagery: Coarse-to-fine text-guided token pruning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 9206–9217, October 2025.

Xianzhi Ma, Shujun Wang, Xiaohan Li, Hao Liu, Changhua Pei, et al. Weaveearth: Structured evidence construction and reasoning for training-free uhr remote sensing understanding. *arXiv preprint arXiv:2607.10120*, 2026.

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.

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. *arXiv preprint arXiv: 2409.19256*, 2024.

Fengxiang Wang, Mingshuo Chen, Yueying Li, Di Wang, Haotian Wang, Zonghao Guo, Zefan Wang, Boqi Shan, Long Lan, Yulin Wang, Hongzhen Wang, Wenjing Yang, Bo Du, and Jing Zhang. Geollava-8k: Scaling remote-sensing multimodal large language models to 8k resolution. In *Advances in Neural Information Processing Systems(NeurIPS)*, 2025a.

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

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*, 2026.

Qianhao Yuan, Jie Lou, Xing Yu, Hongyu Lin, Le Sun, Xianpei Han, and Yaojie Lu. Vision-opd: Learning to see fine details for multimodal llms via on-policy self-distillation. *arXiv preprint arXiv:2605.18740*, 2026.

Yi-Fan 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 *Proceedings of the International Conference on Learning Representations (ICLR)*, 2025.

Siyan Zhao, Zhihui Xie, Mengchen Liu, Jing Huang, Guan Pang, Feiyu Chen, and Aditya Grover. Self-distilled reasoner: On-policy self-distillation for large language models. *arXiv preprint arXiv:2601.18734*, 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.

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

A ALGORITHM OF RS-OPSD

Algorithm 1 summarizes the training procedure of RS-OPSD. For each sample, the student generates an on-policy response from the Global Evidence, while the privileged teacher evaluates the same trajectory using CPVP. CAD then filters privileged supervision according to teacher reliability and student-answer correctness. The student is optimized with the CAD objective and reference-model regularization, while the teacher is updated by EMA.

Algorithm 1 Training procedure of RS-OPSD

Require: Training set $\mathcal{D} = \{(I_G, I_C, I_F, q, a^*)\}$; student π_θ; teacher π_{θ_T}; frozen reference π_{ref}; KL coefficient λ_{KL}; gradient clipping threshold c_{grad}; EMA coefficient β

Initialize $\theta_T \leftarrow \theta$
Freeze π_{ref}
for each training batch $\mathcal{B} \subset \mathcal{D}$ **do**
  for each $(I_G, I_C, I_F, q, a^*) \in \mathcal{B}$ **do**
    Construct privileged information $z \leftarrow \{I_C, I_F\}$
    Sample one student trajectory $\mathbf{y} \sim \pi_\theta(\cdot \mid I_G, q)$
    Construct the valid-answer mask m^{ans} from $\mathbf{y}$
    Compute student token probabilities $p_t^S \leftarrow \pi_\theta(y_t \mid I_G, q, \mathbf{y}_{<t})$
    Compute privileged-teacher probabilities $p_t^T \leftarrow \pi_{\theta_T}(y_t \mid I_G, z, q, \mathbf{y}_{<t})$
    Run the teacher-forced GT-prefix reliability probe: $g \leftarrow \text{Probe}_T(a^*) \in \{0, 1\}$
    Verify the student response: $R \leftarrow +1$ if $\hat{a} = a^*$, otherwise $R \leftarrow -1$
    for each valid answer token y_t **do**
      $r_t \leftarrow \text{sg}[\log p_t^T - \log p_t^S]$
      $A_t \leftarrow g\, R\, [R r_t]_+$
    end for
  end for
  Compute CAD loss

$$\mathcal{L}_{\text{CAD}} = -\frac{\sum_{i,t} m_{i,t}^{\text{ans}}\ \text{sg}[A_{i,t}] \log p_{i,t}^S}{\sum_{i,t} m_{i,t}^{\text{ans}}}$$

  Compute reference-model regularization $\mathcal{L}_{\text{KL}}$
  $\mathcal{L} \leftarrow \mathcal{L}_{\text{CAD}} + \lambda_{\text{KL}} \mathcal{L}_{\text{KL}}$
  Backpropagate $\nabla_\theta \mathcal{L}$
  Clip the student gradient norm with threshold c_{grad}
  Update student parameters θ
  Update teacher by EMA: $\theta_T \leftarrow \beta\theta_T + (1 - \beta)\theta$
end for

For RS-OPD-*Lite*, the same training procedure is used with a Qwen3-VL-2B student and a Qwen3-VL-8B privileged teacher. The teacher remains frozen throughout training.

B DATASET QUALITY AND INTEGRITY

We verified that GeoEvidence-6K has no exact image-level overlap with XLRS-Bench (Wang et al., 2025b), MME-RealWorld-RS (Zhang et al., 2025), or LRS-VQA (Luo et al., 2025), ruling out direct train–test leakage from reused evaluation images. The annotation process involved ten domain experts with expertise in remote sensing image interpretation. All generated samples were manually reviewed, and the complete dataset underwent a final cross-validation stage across the expert pool to resolve remaining ambiguities and annotation errors. We measure annotation reliability using the Cross-Validation Acceptance Rate (CVAR), defined as the proportion of samples accepted during the final cross-validation without requiring further correction. GeoEvidence-6K achieves an overall CVAR of **97.8%**, indicating a high level of consistency in the final annotations.

Figure 6 further summarizes several statistics of GeoEvidence-6K, including the question vocabulary, token-length distribution, and task distribution.

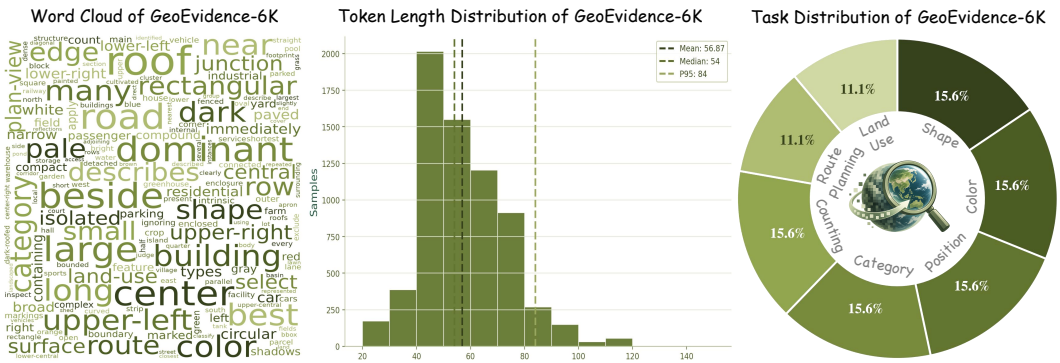


Figure 6: Dataset statistics of GeoEvidence-6K. Left: word cloud of the question texts. Middle: token-length distribution of the questions, with the mean, median, and 95th percentile marked by dashed lines. Right: distribution of task categories.

As shown in Figure 6(left), the most frequent words include terms such as *roof*, *parcel*, *road*, *color*, *shape*, *center*, and relative-location expressions such as *upper-right*, *beside*, and *near*. This indicates that GeoEvidence-6K covers diverse UHR remote sensing VQA patterns, including attribute recognition, object counting, spatial localization, and fine-grained geometric reasoning.

Figure 6(middle) shows that the token-length distribution is concentrated in a moderate range, with a mean of 56.87, a median of 54, and a 95th percentile of 84. Most samples therefore remain relatively compact, while a smaller fraction of longer questions reflects the presence of more compositional and spatially specific instructions.

Figure 6(right) shows that the task categories are overall well balanced. Five local-evidence tasks—*counting*, *position*, *category*, *color*, and *shape*—each account for about 15.6% of the dataset, while the two global-context tasks—*route planning* and *land use*—each account for about 11.1%. This composition provides balanced supervision across both local fine-grained perception and global scene-level understanding.

C Implementation Details

Training framework. We implement RS-OPSD using the `verl` (Sheng et al., 2024) training framework. For each training sample, the student performs a single on-policy rollout ($n = 1$), which is then reused for privileged-teacher evaluation, CAD computation, and reference-model regularization.

Visual preprocessing. During training, all visual inputs are resized while preserving the original aspect ratio, with the longest image side capped at 2,048 pixels. The same preprocessing strategy is applied to the Global Evidence, Contextual Evidence, and Fine-grained Evidence. For training samples with localized evidence annotations, the student Global Evidence contains a rendered red bounding box indicating the annotated evidence region, while the privileged teacher additionally receives the corresponding contextual and fine-grained views. At inference time, we do not introduce any additional bounding boxes, crop coordinates, or localization cues beyond those originally provided by the benchmark. Specifically, MME-RealWorld-RS and LRS-VQA are evaluated using their original unmarked images, whereas a subset of XLRS-Bench samples natively contains red-circle annotations in the released benchmark images.

Training prompts. To ensure reproducibility, Figure 7 summarizes the prompt templates used during RS-OPSD training. The student receives a single full-image input together with the original question, options, and bounding-box instruction when applicable. In contrast, the privileged teacher receives three complementary views—Global Evidence, Contextual Evidence, and Fine-grained Evidence—with an explicit instruction describing the role of each view. The question and

Prompt for the student model with the full-image input.

<image>
{question}
{options}
Only focus on the objects inside the red bounding box in the image to answer this question.

Prompt for the privileged teacher model with Global Evidence, Contextual Evidence, and Fine-grained Evidence as inputs.

<image> \n <image> \n <image>
{question}
{options}
You are given three views of the same aerial scene.
Image 1 is the complete scene for global context.
Image 2 is an expanded high-resolution crop centred on the target; its red bounding box marks the exact target region.
Image 3 is a tight high-resolution crop of that target without a red bounding box.
Use image 3 for the finest target detail, image 2 to identify the exact boxed target and nearby context, and image 1 only when global context is needed.

Figure 7: Prompt templates for the student and privileged teacher in RS-OPSD.

answer options are otherwise kept consistent between the student and teacher prompts. All prompts are wrapped with the Qwen chat template before being passed to the model.

Inference efficiency. Inference latency is measured on a single NVIDIA H100 GPU with a batch size of 1. We report the average per-sample latency under the same evaluation protocol used for the three UHR remote sensing VQA benchmarks. Since RS-OPSD introduces no architectural modifications to the underlying vision-language model and requires neither additional visual search nor external tool calls at inference time, the trained model remains directly compatible with standard inference acceleration frameworks such as vLLM (Kwon et al., 2023), without requiring specialized serving infrastructure.

D Detailed Quantitative Comparison Results

Tables 5 and 6 provide fine-grained quantitative comparison results across the three UHR remote sensing VQA benchmarks.

XLRS-Bench (Wang et al., 2025b). XLRS-Bench evaluates eight perception categories—Overall Counting (OC), Regional Counting (RC), Overall Land Use Classification (OLUC), Regional Land Use Classification (RLUC), Object Classification (OCC), Object Color (OCL), Object Motion State (OMS), and Object Spatial Relationship (OSR)—and five reasoning categories, including Anomaly Detection (AD), Environmental Conditional Reasoning (ECR), Route Planning (RP), Regional Counting with Change Detection (RCCD), and Counting with Complex Reasoning (CCR). As shown in Table 5, RS-OPSD achieves the best overall score of 53.1, exceeding the previous best GeoLLaVA-8K (Wang et al., 2025a) by 1.6 points. It ranks first on RC, RLUC, OCC, ECR, and RCCD, while obtaining the second-best results on OCL and OMS. The improvements span both perception and reasoning tasks, indicating that the gains are not concentrated in a single capability.

MME-RealWorld-RS (Zhang et al., 2025) and LRS-VQA (Luo et al., 2025). As shown in Table 6, RS-OPSD achieves 61.5 on MME-RealWorld-RS, outperforming the previous best Zoom-Search (Zhou et al., 2025) by 3.9 points. In particular, it achieves the best results on Position (76.8) and Color (70.4), while remaining competitive on Count. On LRS-VQA, RS-OPSD reaches the best average score of 33.3, improving over WeaveEarth (Ma et al., 2026) by 2.0 points. Although it does not dominate every individual subset, it achieves consistently strong results across FAIR, Bridge, and STAR, leading to the highest overall average.

Table 5: Detailed quantitative results on XLRS-Bench across perception and reasoning sub-tasks. Green, yellow, and blue denote the first-, second-, and third-best results, respectively.

Method	**Perception**								**Reasoning**					**Avg.**
Sub-tasks	**OC**	**RC**	**OLUC**	**RLUC**	**OCC**	**OCL**	**OMS**	**OSR**	**AD**	**ECR**	**RP**	**RCCD**	**CCR**	
Closed-Source Vision Language Models														
GPT-4o	25.0	32.0	15.0	66.0	9.5	11.3	11.7	24.6	73.0	73.0	35.0	20.0	25.0	32.4
Claude 3.7 Sonnet	27.6	22.7	17.4	68.4	30.5	29.9	63.6	27.6	64.8	78.4	34.5	27.8	32.6	40.5
Gemini 2.5 Pro	**50.0**	42.0	12.0	65.5	37.5	38.1	60.0	30.2	68.0	72.0	36.0	26.7	35.0	44.1
Open-Source Vision Language Models														
LLaVA-OV-7B	25.0	38.0	8.0	69.5	35.9	35.3	65.0	25.2	**76.0**	**83.0**	24.0	43.3	36.0	43.4
Qwen2.5-VL-7B	33.3	40.0	31.0	77.0	40.6	**40.5**	**66.7**	**36.2**	68.0	72.0	27.0	38.3	45.0	47.4
Qwen3-VL-8B	23.3	**45.0**	20.0	**82.0**	**45.9**	**44.4**	**66.7**	30.6	**74.0**	79.0	42.0	**48.3**	**55.0**	**50.5**
InternVL3-8B	**40.0**	39.0	10.0	71.5	**44.5**	30.8	65.0	25.2	**77.0**	**82.0**	36.0	21.7	50.0	45.6
Remote Sensing Vision Language Models														
GeoChat	16.7	29.0	2.0	23.0	21.1	16.8	35.0	24.2	33.0	43.0	10.0	–	21.0	22.9
VHM	18.3	33.0	5.0	38.5	27.5	25.4	35.0	33.0	53.0	58.0	39.0	26.7	34.0	32.8
Specialized Methods for UHR Remote Sensing VQA														
GeoLLaVA-8K	26.7	38.0	**49.0**	69.0	41.6	31.6	65.0	**35.0**	67.0	78.0	**66.0**	**50.0**	**52.0**	**51.5**
UHR-BAT	21.7	33.0	**50.0**	55.5	43.5	33.8	65.0	**44.8**	62.0	71.0	**54.0**	46.7	**51.0**	48.6
ZoomSearch	**46.7**	42.0	12.0	70.5	**44.5**	35.8	**75.0**	32.2	**74.0**	**82.0**	29.0	43.3	37.0	48.0
WeaveEarth	20.0	**47.0**	24.0	**79.0**	43.1	38.0	**66.7**	26.8	69.0	81.0	**46.0**	**48.3**	**55.0**	49.5
ZoomEarth	18.3	35.0	30.0	61.5	39.2	32.4	31.7	20.8	70.0	75.0	38.0	28.3	30.0	39.2
Ours														
RS-OPD-*Lite*	25.0	**45.0**	30.0	72.0	31.1	36.8	63.3	28.0	**74.0**	**83.0**	20.0	36.7	50.0	45.8
RS-OPSD	28.3	**55.0**	**33.0**	**85.0**	**46.8**	**41.9**	**71.7**	29.2	**74.0**	**84.0**	45.0	**51.7**	45.0	**53.1**

Table 6: Detailed quantitative results on MME-RealWorld-RS, LRS-VQA, and XLRS-Bench. Average inference speed is measured in s/sample.

Method	**MME-RealWorld-RS**				**LRS-VQA**				**XLRS-Bench**	**Total**	
Sub-set	**Position**	**Color**	**Count**	**Avg.**	**FAIR**	**Bridge**	**STAR**	**Avg.**	**Avg.**	**Accuracy**	**Speed**
Closed-Source Vision Language Models											
GPT-4o	36.4	32.4	15.9	28.2	22.2	31.8	27.4	27.1	32.4	29.2	–
Claude 3.7 Sonnet	56.6	52.1	28.5	45.7	25.7	29.5	30.1	28.4	40.5	38.2	–
Gemini 2.5 Pro	60.6	60.3	34.3	51.7	**28.9**	31.2	32.7	**30.9**	44.1	42.2	–
Open-Source Vision Language Models											
LLaVA-OV-7B	64.8	61.4	33.8	53.3	20.6	**35.1**	26.1	27.3	43.4	41.3	2.38
Qwen2.5-VL-7B	55.9	48.5	18.6	41.0	23.9	34.2	30.2	29.4	47.4	39.3	1.87
Qwen3-VL-8B	56.4	48.1	21.3	41.9	24.6	**38.2**	27.5	30.1	**50.5**	40.8	1.75
InternVL3-8B	**71.0**	**62.6**	31.4	55.0	24.8	33.4	29.3	29.2	45.6	43.3	**1.25**
Remote Sensing Vision Language Models											
GeoChat	25.1	23.1	15.7	21.3	20.2	24.5	13.8	19.5	22.9	21.2	1.33
VHM	35.2	20.3	16.8	24.1	24.3	27.5	28.3	26.7	32.8	27.9	2.64
Specialized Methods for UHR Remote Sensing VQA											
GeoLLaVA-8K	34.9	27.9	22.3	28.4	21.8	29.2	26.4	25.8	**51.5**	35.2	**1.30**
UHR-BAT	44.0	42.0	14.0	33.3	16.6	23.5	20.0	20.0	48.6	34.0	7.40
ZoomSearch	67.6	**66.1**	**39.2**	**57.6**	25.9	31.2	33.5	30.2	48.0	**45.3**	43.5
WeaveEarth	51.4	27.7	**62.6**	47.2	**31.0**	26.1	**36.7**	**31.3**	49.5	42.7	2.94
ZoomEarth	46.6	41.9	27.6	38.7	20.0	23.2	21.6	21.6	39.2	33.2	14.2
Ours											
RS-OPD-*Lite*	**74.2**	61.9	32.5	**56.2**	27.9	29.6	**33.9**	30.5	45.8	**44.2**	**1.09**
RS-OPSD	**76.8**	**70.4**	**37.4**	**61.5**	**29.4**	**34.9**	**35.6**	**33.3**	**53.1**	**49.3**	1.58

Overall accuracy and efficiency. Averaged across the three benchmarks, RS-OPSD reaches 49.3, a 4.0-point improvement over the strongest competing method. Meanwhile, RS-OPD-*Lite* achieves an average accuracy of 44.2 with only a 2B student, surpassing all evaluated 8B-scale baselines. It also achieves the lowest average inference time at 1.09 s/sample, compared with 1.25 s/sample for the second-fastest method. The full RS-OPSD remains efficient at 1.58 s/sample, even faster than its Qwen3-VL-8B (Bai et al., 2025a) base model at 1.75 s/sample. These results further show that internalizing privileged visual information through OPSD provides a favorable balance between accuracy and inference efficiency.

E Limitations and Future Work

The current formulation of RS-OPSD relies on explicit supervision at two stages. First, target-region annotations are required to construct the privileged visual inputs used by CPVP. Second, ground-truth answers are used in CAD to assess teacher reliability and determine the correctness of student rollouts. Although these annotations enable reliable privileged distillation, acquiring them for UHR remote sensing imagery introduces substantial annotation cost and limits the scalability of the framework.

A natural extension is to explore more self-supervised forms of OPSD that reduce or remove these dependencies. Future work may investigate automatic discovery of question-relevant visual privilege without target-region annotations, as well as teacher reliability estimation without ground-truth answer constraints. More broadly, combining model-intrinsic evidence discovery with self-verification could enable privileged self-distillation to scale to larger unlabeled UHR datasets and broader high-resolution vision-language tasks.

Our RS-OPD-*Lite* experiments further suggest that visual privilege remains beneficial beyond the self-distillation setting, where a larger privileged teacher transfers knowledge to a smaller student. This observation opens another direction for *privilege-enhanced model distillation*, where additional training-time information can improve knowledge transfer across model scales while retaining a lightweight student at inference time. Exploring such privileged distillation may provide a promising route toward more accurate and efficient vision-language models.