GeoOPD: Internalizing Visual Search from Privileged Evidence for Remote Sensing Visual Reasoning

Anonymous authors
Paper under double-blind review

Abstract

Applying multimodal large language models (MLLMs) to remote sensing visual reasoning remains challenging because task-relevant evidence is often fine-grained and sparse. Such evidence can be overwhelmed by extensive background content, leading to incomplete perception and visually unsupported answers. Recent "thinking with images" methods expose local details through iterative cropping, but incur additional inference cost and remain vulnerable to redundant or inaccurate region selection; local views are also insufficient for questions that depend on global context. Motivated by this observation, we introduce GeoOPD, a remote sensing post-training framework that uses Evidence-Privileged On-Policy Distillation (E-OPD) to transfer the benefits of training-time evidence localization into tool-free visual reasoning at deployment. We develop an evidence-driven data synthesis pipeline that constructs marked views and textual evidence privileges according to the spatial organization required by each question. During post-training, the teacher additionally observes the organized evidence and provides token-level distributional supervision over student-generated trajectories. Divergence gating further concentrates the learning signal on response positions where privileged evidence induces substantial predictive changes. Experiments across multiple benchmarks show that GeoOPD consistently improves its backbone and performs favorably against both general-purpose and remote-sensing-specific models. These results suggest that task-relevant evidence available only during training can be partially internalized by a model used for standard inference, complementing iterative visual tool use.

1 Introduction

Multimodal large language models (MLLMs) have made substantial progress in visual understanding and reasoning (Singh et al., 2025; Comanici et al., 2025), and have recently been extended to remote sensing visual question answering, scene understanding, and geospatial reasoning (Kuckreja et al., 2024; Cai et al., 2026). Remote sensing images, however, cover extensive geographic areas in which task-relevant evidence is often fine-grained, sparse, and easily overwhelmed by background content. This can result in incomplete perception and visually unsupported answers, making it difficult to identify relevant evidence without losing the broader scene context.

Some studies address this challenge through the "thinking with images" paradigm, in which models iteratively crop and inspect candidate regions during inference (Zheng et al., 2026; Zhang et al., 2025a). These operations can expose local details that are difficult to perceive in the full scene, but depend heavily on accurate region selection. Incorrect crops may discard useful context or introduce irrelevant observations, while multiple tool calls and image re-encoding increase inference cost (Wang et al., 2026). Moreover, some questions require global context or joint reasoning over multiple spatially distributed objects and therefore cannot be solved from an isolated local view. Recent analyses further suggest that gains from visual tool-use reinforcement learning may partly arise from improved intrinsic model capabilities and fewer tool-induced errors rather than more effective tool use itself (Ma et al., 2026; Wei et al., 2026). This motivates a fundamental question: *can visual search learned from additional training-time evidence be internalized into a model that reasons from standard inputs without iterative visual tools?*

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

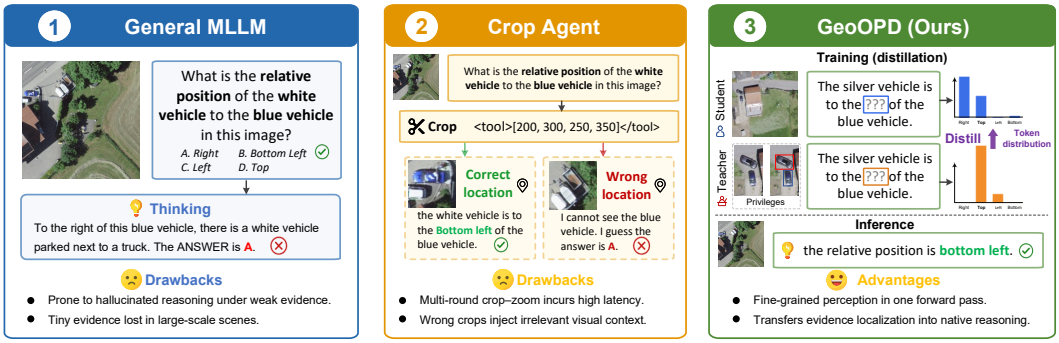


Figure 1: **Comparison of visual reasoning paradigms.** General MLLMs may overlook small evidence in large scenes. Crop agents improve local visibility through iterative operations but introduce additional latency and region-selection errors. GeoOPD transfers evidence-guided teacher corrections during training and performs tool-free inference at deployment.

On-policy self-distillation (OPSD) provides a natural basis for addressing this question (Zhao et al., 2026). A student generates its own response trajectory, while a teacher with access to training-only privileged context evaluates the same prefixes and provides token-level distributional supervision (Yuan et al., 2026). The central challenge is determining what information should constitute this privileged context. A single localized crop cannot simultaneously support local attributes, multi-object relations, and global scene understanding. Conversely, a gold answer can reduce teacher uncertainty but may introduce answer-conditioned shortcuts without improving visual grounding. Effective teacher privileges should therefore expose task-relevant evidence while preserving the context required to interpret it.

We introduce **GeoOPD**, a post-training framework for remote sensing centered on **E**vidence-Privileged **O**n-**P**olicy Self-**D**istillation (E-OPD). GeoOPD organizes teacher evidence into five forms: single-object, paired-object, regional, instance-set, and global evidence, according to the spatial scope required by each question. An evidence-driven synthesis pipeline constructs question–answer (QA) samples together with marked views, focused visual observations, and textual evidence descriptions. Multimodal verification and rule-based quality control remove unsupported, ambiguous, and internally inconsistent candidate samples. During E-OPD, the student receives only the deployment image and question, while a same-checkpoint teacher additionally observes the organized evidence and provides token-level supervision over student-generated trajectories. GeoOPD further employs divergence gating to focus optimization on response positions where privileged evidence induces meaningful predictive changes. Figure 1 compares GeoOPD with standard MLLMs and inference-time crop agents.

We instantiate the framework as GeoOPD-4B and evaluate it on in-domain and zero-shot remote sensing benchmarks. GeoOPD consistently improves its backbone and performs favorably against general-purpose and remote-sensing-specific models without inference-time tool calls. Ablations further demonstrate the effects of post-training stages, teacher privileges, and divergence-based token selection.

Our main contributions are summarized as follows:

- We formulate tool-free remote sensing visual reasoning as the transfer of training-time privileged evidence and propose **GeoOPD**, whose core E-OPD objective distills evidence-induced teacher corrections along student-generated trajectories.
- We develop an evidence-driven data synthesis pipeline that converts both unannotated images and existing question–answer pairs into grounded training samples. The supporting evidence is verified and organized into five spatial forms to provide visual and textual teacher privileges.
- We introduce divergence-gated distillation to focus supervision on privilege-sensitive response positions. Experiments on in-domain and zero-shot benchmarks demonstrate consistent improvements.

2 Related Works

Geospatial Vision-Language Reasoning. Early geospatial vision-language models, such as GeoChat (Kuckreja et al., 2024) and SkyEyeGPT (Zhan et al., 2025), mainly focused on visual question answering and region-level understanding. GeoCoT (Li et al., 2026) subsequently introduced chain-of-thought reasoning to enable structured, step-by-step analysis of remote sensing imagery, while GeoViS (Zhang et al., 2026b) further formulated visual grounding as a progressive search-and-reasoning process. More recently, DeepEyes (Zheng et al., 2026) introduced "thinking-with-images" approaches that incorporate image cropping and zooming into the reasoning process, enabling active perception by acquiring localized visual evidence. While DeepEyes explored this paradigm in the general domain, ZoomEarth (Liu et al., 2026b) adapted active perception to ultra-high-resolution remote sensing with domain-specific training on a dedicated dataset. However, such iterative perception incurs additional inference latency and contextual accumulation, while redundant or uninformative views may provide limited evidence or even interfere with reasoning when the global context is sufficient.

On-Policy Self-Distillation. On-policy self-distillation (Zhao et al., 2026) applies on-policy distillation (Agarwal et al., 2024) in a self-distillation setting, where a single model serves as both teacher and student under different conditioning contexts. Early work explored OPSD for mathematical reasoning in large language models (Shen et al., 2026). Vision-OPD (Yuan et al., 2026) then extended OPSD to multimodal vision-language models by using privileged regional visual information to supervise the full-image model. Subsequent works further investigated the quality and selectivity of visual supervision. VAD (Zhang et al., 2026a) focused on attributing teacher corrections to visual evidence through counterfactual target reconstruction, while OPD-V (Bi et al., 2026) introduced modality-aware token selection to mitigate modality imbalance during self-distillation. However, these methods mainly rely on local or fixed visual interventions, which may be insufficient for remote sensing tasks that require global spatial context. Moreover, sequence-level supervision may introduce optimization noise by emphasizing tokens that are less dependent on visual evidence.

3 Preliminaries

On-Policy Self-Distillation. OPSD samples a response $y = (y_1, \ldots, y_T)$ from the current student policy and uses a frozen teacher to evaluate the same student-generated prefixes under training-only privileged context (Zhao et al., 2026). The teacher shares the student's architecture and is initialized from the same checkpoint, but only the student is updated during distillation. Let x denote the deployment-time input, h the paired privileged context, and $\bar{\theta}$ the fixed teacher parameters. At token position t,

$$\begin{aligned} p_t^S(\cdot) &:= p_\theta(\cdot \mid x, y_{<t}), \\ p_t^T(\cdot) &:= p_{\bar{\theta}}(\cdot \mid x, h, y_{<t}), \qquad y \sim p_\theta(\cdot \mid x). \end{aligned} \tag{1}$$

The student is optimized by matching the token-level teacher distribution along its own sampled trajectory:

$$\mathcal{L}_{\mathrm{OPSD}} = \mathbb{E}_{(x,h)\sim\mathcal{D},\, y\sim p_\theta(\cdot|x)} \left[\frac{1}{T} \sum_{t=1}^{T} D_{\mathrm{dist}} \left(\mathrm{sg}[p_t^T], p_t^S \right) \right], \tag{2}$$

where D_{dist} is a distributional divergence and sg blocks gradients through the teacher. OPSD specifies how privileged supervision is transferred along student-generated trajectories, but leaves privilege construction and informative token selection unspecified.

4 GeoOPD

We propose **GeoOPD**, a post-training framework for remote sensing MLLMs centered on Evidence-Privileged On-Policy Distillation. GeoOPD constructs evidence-grounded training records and organizes their verified evidence into visual and textual teacher privileges. Following supervised fine-tuning (SFT) initialization, E-OPD transfers evidence-induced teacher corrections along student-generated trajectories, while divergence gating selects positions that are most sensitive to the privileged evidence.

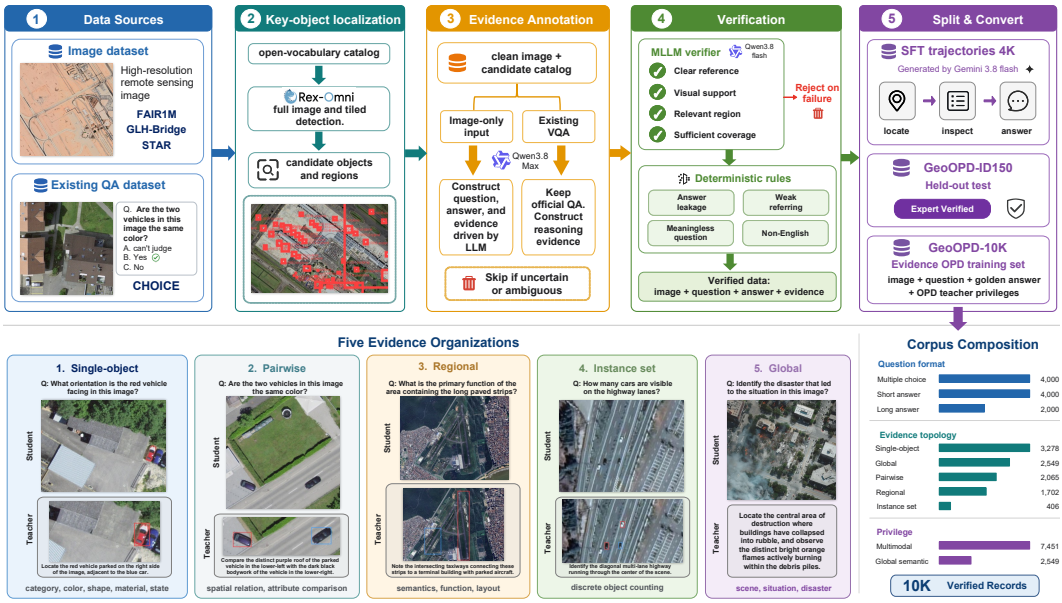


Figure 2: **GeoOPD data synthesis and evidence organization.** The pipeline converts image-only data and existing VQA samples into quality-controlled image–question–answer–evidence records through key-object localization, source-specific annotation, multimodal verification, and rule-based filtering. The resulting records are used to construct SFT trajectories, the held-out GeoOPD-ID150 evaluation set, and organized teacher privileges for E-OPD. The bottom row illustrates the five evidence organizations.

4.1 Data Synthesis Pipeline

High-quality remote sensing VQA data suitable for post-training remain scarce, particularly in terms of task diversity, annotation reliability, and explicit spatial evidence. We therefore develop a synthesis agent that coordinates an open-vocabulary detector with MLLMs, as illustrated in Fig. 2. The training corpus combines LRS-GRO (Liu et al., 2026b) and CHOICE (An et al., 2026). LRS-GRO aggregates high-resolution imagery from FAIR1M (Sun et al., 2022), GLH-Bridge (Li et al., 2024), and STAR (Li et al., 2025), from which we use only the images. CHOICE (An et al., 2026) provides existing image–question–answer triples; we retain unambiguous questions whose answers can be objectively verified from the corresponding image.

We construct an open-vocabulary inventory of geospatial objects and use Rex-Omni (Jiang et al., 2026) to retrieve candidate objects and regions at global and local scales. For dense or visually similar targets, we additionally crop a localized view that serves as the student-facing question image and provides an unambiguous referent. Qwen3.8-Max then examines this image together with the candidate proposals. For LRS-GRO, it selects visually verifiable content and constructs a question, its answer, and the supporting cues. For CHOICE, it preserves the original question and answer while grounding the objects, regions, or scene-level cues that support the answer. A candidate is skipped when visual details are insufficient, the target cannot be uniquely identified, or the available evidence does not support a definitive answer.

An independent Qwen3.8-flash verifier re-examines each candidate against the image, checking referential clarity, visual support for the answer, evidence relevance, and evidence coverage. Candidates that fail any criterion are discarded. Rule-based filters further remove duplicated questions, invalid references, residual internal proposal identifiers, incomplete counting evidence, and student-facing questions whose stems directly reveal their answers. This process produces quality-controlled question–answer pairs with associated evidence records.

The verified records are then assigned according to their training purpose. For SFT initialization, we select 4,000 samples and use Gemini3.8 flash to generate evidence-constrained visual reasoning trajectories. Each trajectory follows the progression of locating the relevant content, examining

decisive visual characteristics, and deriving the answer. A consistency check removes statements unsupported by the image evidence. Samples assigned to E-OPD retain their questions, answers, and spatial evidence and proceed to teacher-privilege construction. We additionally construct GeoOPD-ID150 (ID150) for in-domain evaluation. This training-disjoint, expert-reviewed set contains 75 multiple-choice questions, 50 short-answer questions, and 25 long-answer questions. Further details are provided in Appendix A.3.2.

4.2 Evidence-Organized Privilege Synthesis

Privileged distillation is effective only when the additional teacher context reveals relevant evidence without removing information required by the question. Recent visual OPD methods commonly instantiate privileged information as a single region of interest (Yuan et al., 2026). Although such a view may improve local attribute recognition, it can remove the shared coordinate frame required for spatial relations, omit dispersed instances in counting tasks, or provide little benefit for global scene understanding. We therefore assign each verified sample to one of five evidence organizations according to the spatial extent, cardinality, and relationships of its supporting evidence. Each organization determines how the visual and textual teacher privileges are constructed.

- **Single-object evidence** $S_{\mathbf{object}}$ supports questions about object category, color, shape, material, and state. A contextual marker identifies the target in the scene, a focused view exposes local details, and the textual description records its spatial anchor and visible attributes.
- **Paired-object evidence** $M_{\mathbf{pair}}$ supports spatial relations and attribute comparisons. A joint view preserves both referenced objects and their shared coordinate frame, while the text identifies the objects and their comparison-relevant properties.
- **Regional evidence** $S_{\mathbf{region}}$ supports regional semantics, local functions, and layout understanding. The privilege retains the complete semantic region and describes its boundary, constituent elements, and internal organization.
- **Instance-set evidence** $M_{\mathbf{set}}$ supports counting questions. All relevant instances are jointly marked within the specified scope, and the textual description records the counting boundary and instance-identification criteria.
- **Global evidence** G supports scene classification, situation inference, and disaster assessment. The complete scene is retained without an additional local crop, while a scene-level description summarizes the principal geographic entities, environmental conditions, and global layout.

We use Qwen3.8-flash to convert the verified evidence into concise descriptions of observable visual facts. For single-object, paired-object, regional, and instance-set evidence, these descriptions are combined with marked or focused views to form multimodal privileges. Global evidence primarily uses a scene-level description grounded in the complete image.

The resulting GeoOPD-10K privileged training corpus contains 6,000 open-ended VQA examples, including 4,000 short-answer and 2,000 long-answer questions, together with 4,000 multiple-choice questions. It covers five evidence organizations and 16 fine-grained task types. Among the 10,000 quality-controlled records, 7,451 contain visual privileges and 2,549 use scene-level semantic privileges. The corpus composition and evidence organizations are summarized in Fig. 2.

4.3 Training Procedure

GeoOPD follows a two-stage training procedure. SFT first initializes the model with evidence-constrained remote sensing reasoning trajectories. E-OPD then transfers predictive corrections induced by the organized teacher evidence to the student.

Supervised cold start. We first fine-tune the backbone on the SFT subset described in Section 4.1. The model receives only the student-facing image and original question, and learns to locate relevant content, examine discriminative visual cues, and produce a structured response. The resulting SFT checkpoint initializes both the trainable student and the fixed teacher used in the subsequent distillation stage.

Evidence-privileged on-policy distillation. Let a training example be $z = (I, q, E, o)$, where I is the student-facing image, q is the question, E is the verified evidence, and o denotes its evidence organization. The student receives the deployment input $x_z = (I, q)$, while the teacher additionally receives

$$h_o = \mathcal{C}_o(I, E), \tag{3}$$

where $\mathcal{C}_o$ denotes the privilege construction procedure described in Section 4.2. All evidence organizations share the same teacher; only the privileged context h_o changes across examples.

Following the OPSD formulation in Section 3, the student samples a trajectory from the deployment input, while the teacher evaluates the same student-generated prefixes with the additional evidence. We instantiate D_{dist} in Eq. equation 2 using the symmetric Jensen-Shannon divergence. At response position t, let

$$\begin{aligned} \mu_t &= \frac{1}{2}\left(p_t^S + \text{sg}[p_t^T]\right), \\ d_t &= \frac{1}{2} D_{\text{KL}}\left(\text{sg}[p_t^T] \,\middle\|\, \mu_t\right) + \frac{1}{2} D_{\text{KL}}\left(p_t^S \,\middle\|\, \mu_t\right). \end{aligned} \tag{4}$$

Here, d_t measures the token-level discrepancy between the student and the evidence-privileged teacher under the same generated prefix. Gradients are propagated only through the student distribution.

Visual reasoning trajectories have non-uniform information density. Generic transitions and formatting tokens are often insensitive to the teacher evidence, whereas attributes, spatial relations, quantities, and answer decisions tend to produce larger discrepancies. Uniformly averaging over all response positions can therefore dilute the supervision carried by critical tokens (Xu et al., 2026). We address this issue through divergence-gated token distillation.

Let $\mathcal{V}(y)$ denote the valid response positions in trajectory y. For a retention ratio γ, the number of selected positions is

$$K(y) = \min\left(|\mathcal{V}(y)|, \max\left(K_{\min}, \lceil \gamma |\mathcal{V}(y)| \rceil\right)\right). \tag{5}$$

We then retain the positions with the largest token-level divergences:

$$\mathcal{G}_\gamma(y) = \left\{ t \in \mathcal{V}(y) \;\middle|\; \text{rank}_{\mathcal{V}(y)}(d_t) \le K(y) \right\}, \tag{6}$$

where rank one corresponds to the largest divergence. The E-OPD objective is

$$\mathcal{L}_{\text{E-OPD}} = \mathbb{E}_{z \sim \mathcal{D}_E,\, y \sim p_\theta(\cdot|x_z)} \left[\frac{1}{|\mathcal{G}_\gamma(y)|} \sum_{t \in \mathcal{G}_\gamma(y)} d_t \right]. \tag{7}$$

The minimum-token constraint $K_{\min}$ ensures that every valid trajectory receives supervision. Setting $\gamma = 1$ recovers full-sequence E-OPD, while our main setting uses $\gamma = 0.5$ and retains half of each trajectory with the largest teacher–student discrepancies. This selection requires neither an additional negative teacher nor artificially masked images. A small KL penalty to the frozen SFT reference is additionally used to limit policy drift; its coefficient and other optimization details are provided in the Appendix A.2.

5 Experiments

5.1 Experimental Setup

Training and inference. We initialize GeoOPD from Qwen3-VL-4B-Instruct and sequentially perform SFT and E-OPD using the data constructed in Section 4.1. At inference time, GeoOPD receives no privileged information and reasons solely from the input image and question. Proprietary and general-purpose open-source MLLMs use a common prompt that requests a concise rationale followed by the final answer. For remote-sensing-specific MLLMs, we follow their released inference prompts and native tool-execution protocols whenever available.

Benchmarks and baselines. We use ID150 for in-domain evaluation and additionally evaluate on MME-RealWorld-RS (Zhang et al., 2025b) and XLRS-Bench (Wang et al., 2025) to assess zero-shot generalization. Predictions are first processed by deterministic rule-based extractors, and an LLM

Table 1: **Main comparison across remote sensing VQA benchmarks.** We report accuracy (%) on GeoOPD-ID150 (ID150), the remote sensing subset of MME-RealWorld (MME-RealWorld-RS), and XLRS-Bench. Overall is the unweighted mean of the three benchmark-level scores. **Bold** indicates the best result in each column.

Method	ID150			MME-RealWorld-RS				XLRS-bench	Overall
	Choice	Short	Avg.	Pos.	Color	Count	Avg.	All	
Proprietary MLLMs									
GPT-6 Luna	64.0	48.0	57.6	**66.0**	**62.0**	**48.0**	**58.7**	49.2	55.2
Gemini 2.5 Pro	40.0	42.0	40.8	60.0	60.0	38.0	52.7	45.2	46.2
General-purpose open-source MLLMs									
Qwen2.5-VL-7B	52.0	38.0	46.4	52.0	38.0	14.0	34.7	39.0	40.0
Qwen3-VL-4B	60.0	76.0	66.4	60.0	54.0	12.0	42.0	42.4	50.3
Qwen3.5-4B	65.3	80.0	71.2	64.0	52.0	24.0	46.7	21.3	46.4
Remote-sensing-specific MLLMs									
GeoChat-7B	40.0	18.0	31.2	30.0	38.0	10.0	26.0	31.0	29.4
ZoomEarth-3B	44.0	46.0	44.8	40.0	50.0	30.0	40.0	40.2	41.7
RSThinker-9B	60.0	60.0	60.0	44.0	46.0	14.0	34.7	29.5	41.4
GeoOPD (Ours)	**72.0**	**82.0**	**76.0**	62.0	50.0	30.0	47.3	**59.0**	**60.8**

judge is invoked only when rule-based parsing fails. The proprietary baselines are GPT-6 Luna and Gemini 2.5 Pro (Comanici et al., 2025). The parameter-comparable general-purpose baselines include Qwen2.5-VL-7B-Instruct (Bai et al., 2025b), Qwen3-VL-4B-Instruct (Bai et al., 2025a), and Qwen3.5-4B, while the remote-sensing-specific baselines include GeoChat-7B (Kuckreja et al., 2024), ZoomEarth-3B (Liu et al., 2026b), and RSThinker (Liu et al., 2026a).

Implementation details. All training is conducted on 2 NVIDIA H200 GPUs. We run SFT and E-OPD for two and one epochs, respectively. E-OPD retains the top 50% of response positions ranked by token-level divergence. We set the decoding temperature to zero for evaluation. Additional training configurations, prompts, and evaluation details are provided in Appendix A.2.

5.2 Main Results

Table 1 compares GeoOPD with proprietary, general-purpose, and remote-sensing-specific models across three benchmarks.

In-domain performance. GeoOPD achieves 76.0% overall accuracy on ID150, with 72.0% on multiple-choice questions and 82.0% on short-answer questions. Compared with its Qwen3-VL-4B backbone, GeoOPD improves accuracy on the two answer formats by 12.0 and 6.0 percentage points, respectively.

ID150 includes questions about local object attributes, spatial relations, and counting, together with questions that require global scene understanding. Strong performance therefore requires both fine-grained evidence localization and preservation of broader scene context. Remote-sensing-specific models are also competitive on this benchmark: RSThinker reaches 60.0%, slightly exceeding the proprietary baselines, while ZoomEarth obtains 44.8%, compared with 40.8% for Gemini 2.5 Pro. GeoOPD improves both answer formats without inference-time tools, indicating that training-time evidence supervision can improve the model's ability to search the input image for relevant evidence.

Zero-shot performance. Compared with the satellite imagery in GeoOPD-ID150, MME-RealWorld-RS emphasizes fine-grained questions over aerial imagery, providing a stricter test of cross-domain generalization and local evidence search. GeoOPD achieves 47.3%, outperforming Qwen3-VL-4B by 5.3 points. Its accuracies on position, color, and counting questions are 62.0%, 50.0%, and 30.0%, respectively, yielding the strongest aggregate result among the evaluated open-source and remote-sensing-specific models. Despite its compact 4B scale, GeoOPD remains 11.4 points behind GPT-6 Luna, with the largest gap in counting, indicating that enumerating small or spatially dispersed instances remains challenging.

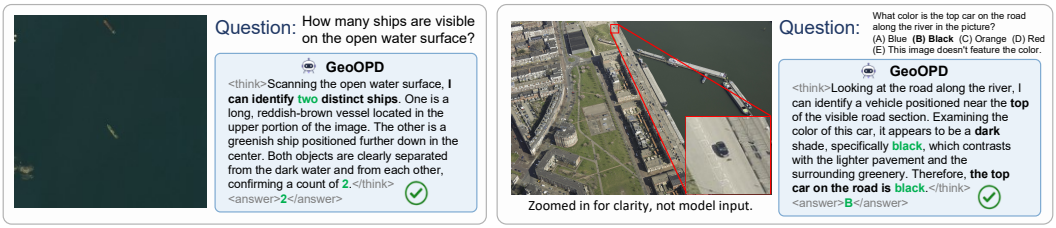


Figure 3: **Qualitative examples of GeoOPD at deployment.** GeoOPD receives only the input image and question. **Left:** it identifies two ships on the open water and answers "2." **Right:** it identifies the uppermost vehicle along the river road as black and selects (B). The red marker and enlarged inset are included only for visualization.

Table 2: Comparison of post-training schedules.

Training schedule	ID150	MME-RS
SFT	72.0	43.3
SFT + GRPO	76.0	46.0
SFT + E-OPD	76.0	47.3
SFT + E-OPD + GRPO	76.8	46.0

Table 3: Teacher privilege ablation.

Teacher privilege	ID150	MME-RS
None (SFT)	72.0	43.3
Evidence	76.0	47.3
Gold	77.6	45.3
Evidence + Gold	76.0	46.7

GeoOPD reaches 59.0% on XLRS-Bench, outperforming Qwen3-VL-4B and GPT-6 Luna by 16.6 and 9.8 points, respectively. This result indicates that the model can use question-provided spatial cues to search for fine-grained targets in large remote sensing scenes.

Qualitative examples. Figure 3 presents two deployment-time examples in which GeoOPD receives only the input image and question. In the GeoOPD-ID150 example, it identifies two low-contrast ships on the open water without mistaking harbor structures near the image boundary for additional instances. In the MME-RealWorld-RS example, it localizes the relevant vehicle along the river road and correctly recognizes its color despite the surrounding roads and buildings. The visual annotations and enlarged region are included only for presentation. These examples illustrate GeoOPD's ability to locate sparse instances and resolve details of small targets.

5.3 Ablation Studies

Post-training stages. To isolate the contribution of E-OPD and compare it with standard reinforcement learning, we evaluate four training schedules: SFT, SFT followed by GRPO, SFT followed by E-OPD, and SFT followed by E-OPD and GRPO. All post-SFT variants start from the same SFT checkpoint and share the same backbone. As shown in Table 2, GRPO improves SFT from 72.0% to 76.0% on GeoOPD-ID150 and from 43.3% to 46.0% on MME-RealWorld-RS. E-OPD reaches the same in-domain accuracy while further improving MME-RealWorld-RS to 47.3%, exceeding GRPO by 1.3 points on the external benchmark. Applying GRPO after E-OPD produces a small in-domain gain but reduces MME-RealWorld-RS to 46.0%, yielding no consistent improvement. In this setting, the dense token-level supervision provided by E-OPD transfers more effectively across datasets than outcome-based policy optimization.

Teacher privilege composition. In OPSD, the learning signal arises from the predictive difference between the student and the privileged teacher. Consequently, the information provided to the teacher directly determines which corrections can be transferred. We compare three teacher conditions: organized visual and textual evidence, the gold answer, and their combination. The SFT checkpoint, which receives no privileged teacher supervision, serves as the reference. Table 3 shows that evidence privilege improves SFT by 4.0 points on both benchmarks, reaching 76.0% on GeoOPD-ID150 and 47.3% on MME-RealWorld-RS. Gold-answer privilege performs best in-domain at 77.6%, but reaches only 45.3% on the external benchmark. Evidence privilege therefore provides the stronger zero-shot result, whereas the gold answer more directly improves the in-domain task. Combining the two privileges does not yield an additional gain, reaching 76.0% and 46.7%. These results suggest that gold answers sharpen task-specific predictions, while organized evidence provides visual corrections that transfer more effectively across datasets.

Table 4: Comparison of divergence-gating strategies.

Token selection	ID150	MME-RS
Full sequence	74.4	46.0
Hard threshold	74.4	46.7
Top-ratio (50%)	76.0	47.3
Soft reweighting	73.6	44.7

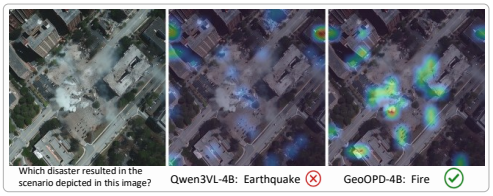


Figure 4: Attention heatmaps of GeoOPD.

Divergence-based token selection. Response tokens differ substantially in their sensitivity to privileged evidence (Xu et al., 2026). Generic transitions and formatting tokens often produce similar student and teacher distributions, whereas attributes, relations, quantities, and answer decisions tend to produce larger discrepancies. We compare four strategies: full-sequence distillation, hard threshold filtering, top-ratio selection, and soft divergence reweighting. As shown in Table 4, retaining the top 50% of response positions ranked by token-level divergence performs best on both benchmarks. Compared with full-sequence distillation, it improves ID150 by 1.6 points and MME-RealWorld-RS by 1.3 points. Hard thresholding provides a smaller improvement, while soft reweighting reduces performance to 73.6% and 44.7%. Thus, per-trajectory relative selection concentrates supervision on privilege-sensitive positions more robustly than hard thresholding or soft weighting.

5.4 Analysis and Discussion

Attention visualization. Attention maps provide an intuitive view of which visual regions a model emphasizes when forming its prediction (Khayatkhoei et al., 2025). Figure 4 visualizes the attention patterns of Qwen3-VL-4B and GeoOPD under the same image–question input. Neither model receives teacher evidence during inference. Qwen3-VL-4B distributes its attention across several background regions and predicts an earthquake, whereas GeoOPD concentrates more strongly on the smoke-affected areas and correctly identifies a fire. This shift is qualitatively consistent with improved selection of task-relevant visual evidence after E-OPD training.

Internalized evidence selection and explicit information acquisition. Our results suggest an important distinction between selecting existing evidence and acquiring new information. If an agent can reliably decide where to crop from its initial visual encoding, that decision already contains partial information about the location of the relevant evidence. In such cases, subsequent cropping may primarily confirm or magnify a region that the model has already identified. E-OPD moves this recoverable part of visual search into post-training, converting recurring evidence-selection patterns into a native model capability. GeoOPD can therefore avoid unnecessary tool calls, reduce region-selection errors, and preserve the full scene context during standard inference.

Active perception remains valuable when a task requires information that is unavailable in the model's current visual representation (Wei et al., 2026). When a question explicitly specifies a region of interest, cropping that region from the original high-resolution image can reveal details that are not preserved in the standard model input. Detection, segmentation, multispectral analysis, and tools equipped with external knowledge can likewise introduce verifiable observations and reduce unsupported inference. GeoOPD and tool-augmented agents are therefore complementary. GeoOPD internalizes evidence selection when the relevant signal is already available, while external tools should be reserved for cases that require new observations or specialized prior knowledge.

6 Conclusion

We introduced GeoOPD, a remote sensing post-training framework based on Evidence-Privileged On-Policy Distillation. It constructs verified multimodal privileges across five evidence organizations and transfers teacher corrections to a student receiving only the standard image–question input. Divergence gating focuses supervision on privilege-sensitive positions. Experiments show consistent in-domain and zero-shot improvements without inference-time tools. More broadly, GeoOPD demonstrates that visual evidence selection can be partially internalized during training, reducing unnecessary interaction while complementing explicit information acquisition. Future work may integrate adaptive detection, segmentation, and multispectral tools to improve performance on complex remote sensing tasks.

AI USE STATEMENT

In this work, we used generative AI tools to assist with language editing and proofreading to improve the clarity, grammar, and readability of the manuscript. We also used generative AI tools to assist with LLM training environment configuration, data format conversion, and the preparation and anonymization of the released code. The use of large language models for synthetic data generation, data filtering, and knowledge distillation is part of our proposed research pipeline rather than AI-assisted authoring.

We did not use generative AI tools to generate the manuscript's scientific content, experimental results, or conclusions. All AI-assisted outputs were reviewed by the authors, and code, training configurations, data-processing procedures, and experimental results were independently verified and tested. We take full responsibility for the final content of this work, including all text, claims, data, results, and artifacts produced with the aid of generative AI.

REFERENCES

Rishabh Agarwal, Nino Vieillard, Yongchao Zhou, Piotr Stanczyk, Sabela Ramos Garea, Matthieu Geist, and Olivier Bachem. On-policy distillation of language models: Learning from self-generated mistakes. In *International Conference on Learning Representations*, volume 2024, pp. 21246–21263, 2024.

Xiao An, Jiaxing Sun, Zihan Gui, and Wei He. Choice: benchmarking the remote sensing capabilities of large vision-language models. *Advances in Neural Information Processing Systems*, 38, 2026.

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, 2025b. URL https://arxiv.org/abs/2502.13923.

Jinhe Bi, Peng Liao, Zengjie Jin, Volker Tresp, Fei Shen, Yunpu Ma, Tat-Seng Chua, et al. Opd-v: Visual on-policy self-distillation with modality balance. *arXiv preprint arXiv:2608.05131*, 2026.

Miaoxin Cai, Guanqun Wang, Wei Zhang, Guangyao Zhou, Yin Zhuang, Tong Zhang, Hao Wang, He Chen, and Jun Li. Earth-onevision: Extending remote sensing multimodal large language models to more sensor modalities and tasks. *arXiv preprint arXiv:2606.10819*, 2026.

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.

Qing Jiang, Junan Huo, Xingyu Chen, Yuda Xiong, Zhaoyang Zeng, Yihao Chen, Tianhe Ren, Junzhi Yu, and Lei Zhang. Detect anything via next point prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 25472–25483, 2026.

Mahyar Khayatkhoei, Prateek Chhikara, Filip Ilievski, et al. Mllms know where to look: Training-free perception of small visual details with multimodal llms. In *International Conference on Learning Representations*, volume 2025, pp. 68194–68213, 2025.

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

Daixun Li, Zirui Li, Sibo He, Jiayun Tian, Mingxiang Cao, Weiying Xie, Yunke Wang, Xin Zhang, Yusi Zhang, Yunsong Li, Chang Xu, and Leyuan Fang. Geocot: Towards reliable remote sensing reasoning with manifold perspective. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 20575–20585, June 2026.

Yansheng Li, Junwei Luo, Yongjun Zhang, Yihua Tan, Jin-Gang Yu, and Song Bai. Learning to holistically detect bridges from large-size vhr remote sensing imagery. *IEEE transactions on pattern analysis and machine intelligence*, 46(12):11507–11523, 2024.

Yansheng Li, Linlin Wang, Tingzhu Wang, Xue Yang, Junwei Luo, Qi Wang, Youming Deng, Wenbin Wang, Xian Sun, Haifeng Li, et al. Star: A first-ever dataset and a large-scale benchmark for scene graph generation in large-size satellite imagery. *IEEE Trans. Pattern Anal. Mach. Intell.*, 47(3):1832–1849, 2025.

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

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

Yan Ma, Weiyu Zhang, Tianle Li, Linge Du, Xuyang Shen, and Pengfei Liu. What does vision tool-use reinforcement learning really learn? disentangling tool-induced and intrinsic effects for crop-and-zoom. *arXiv preprint arXiv:2602.01334*, 2026.

Zhanming Shen, Jintao Tong, Shaotian Yan, Chen Shen, Hao Chen, Wentao Ye, Xiaomeng Hu, Rui Miao, Haobo Wang, Junbo Zhao, et al. Purified opsd: On-policy self-distillation without losing how to think. *arXiv preprint arXiv:2607.02234*, 2026.

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

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

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

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.

Lai Wei, Liangbo He, Jun Lan, Lingzhong Dong, Yutong Cai, Siyuan Li, Huijia Zhu, Weiqiang Wang, Linghe Kong, Yue Wang, et al. Zooming without zooming: Region-to-image distillation for fine-grained multimodal perception. *arXiv preprint arXiv:2602.11858*, 2026.

Yuanda Xu, Hejian Sang, Zhengze Zhou, Ran He, Zhipeng Wang, and Alborz Geramifard. Tip: Token importance in on-policy distillation. *arXiv preprint arXiv:2604.14084*, 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.

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

Kangning Zhang, Yixing Li, Shuai Shao, Qingyao Li, Zhengxi Lu, Zhiyuan Yao, Jianghao Lin, Wenxiang Jiao, Yuan Lu, Weiwen Liu, et al. Vad: Attributing visual evidence for target reconstruction in multimodal on-policy distillation. *arXiv preprint arXiv:2607.28590*, 2026a.

Peirong Zhang, Yidan Zhang, Luxiao Xu, Jinliang Lin, Zonghao Guo, Fengxiang Wang, Xue Yang, Kaiwen Wei, and Lei Wang. Geovis: Geospatially rewarded visual search for remote sensing visual grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14335–14345, 2026b.

Yi-Fan Zhang, Xingyu Lu, Shukang Yin, Chaoyou Fu, Wei Chen, Xiao Hu, Bin Wen, Kaiyu Jiang, Changyi Liu, Tianke Zhang, Haonan Fan, Kaibing Chen, Jiankang Chen, Haojie Ding, Kaiyu Tang, Zhang Zhang, Liang Wang, Fan Yang, Tingting Gao, and Guorui Zhou. Thyme: Think beyond images, 2025a. URL https://arxiv.org/abs/2508.11630.

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.

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.

Yaowei Zheng, Richong Zhang, Junhao Zhang, Yanhan Ye, and Zheyan Luo. Llamafactory: Unified efficient fine-tuning of 100+ language models. In *Proceedings of the 62nd annual meeting of the association for computational linguistics (volume 3: system demonstrations)*, pp. 400–410, 2024.

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

A APPENDIX

A.1 BASELINE MODELS

To evaluate the performance of our approach, we benchmark against three distinct categories of baselines:

Proprietary Models. We include two representative closed-source models to establish proprietary multimodal references. **GPT-6 Luna** (released on September 23, 2026) serves as the latest multimodal LLM that balances high performance with high cost-effectiveness. **Gemini 2.5 Pro** represents a standard commercial baseline widely adopted in vision-language evaluation.

General-Purpose Open-Source Models. We evaluate **Qwen2.5-VL-7B**, **Qwen3-VL-4B**, and **Qwen3.5-4B** to assess general vision-language capabilities. Among these, Qwen3-VL-4B also serves as the foundational backbone for our proposed GeoOPD. All general-purpose baselines operate in a standard single-pass inference mode without external tools or dynamic cropping.

Remote-Sensing Vision-Language Models. To evaluate domain-specific adaptations, we compare with three domain-tailored models: (1) **GeoChat**, an early remote-sensing VLM designed for grounded image understanding; (2) **ZoomEarth**, built upon Qwen2.5-VL-3B, which incorporates dynamic cropping and zooming mechanisms during inference; and (3) **RSThinker**, fine-tuned with remote-sensing chain-of-thought supervision, which generates reasoning paths directly in a single forward pass without physical image manipulations.

A.2 TRAINING AND EVALUATION PROTOCOL

Training runs on one node with two NVIDIA H200 NVL GPUs, each with 141 GB of memory. SFT fully updates the language model in LLaMA-Factory (Zheng et al., 2024) under DeepSpeed ZeRO-3, bfloat16, and FlashAttention-2. The vision tower and multimodal projector remain frozen. Each GPU uses a micro-batch size of 1 and accumulates gradients over four steps; with a tensor-parallel size of 2, so the effective batch size is 8. The learning rate is 1×10^{-5} with a cosine schedule, 30 warmup steps, and weight decay 0.1. SFT runs for two epochs and takes about 4 GPU-hours.

E-OPD continues from the SFT checkpoint for one epoch on the same two GPUs. Parameters are sharded with FSDP, and student rollouts are served by vLLM with tensor parallelism 2. The effective batch size is 8 and the learning rate is 2×10^{-6}. The distillation loss is a symmetric Jensen–Shannon divergence on the top 50% of response tokens ranked by token-level divergence. A KL penalty of 0.02 keeps the student close to the frozen SFT policy. E-OPD takes about 13 GPU-hours.

Local models are decoded deterministically. Each image is limited to 2048×2048 pixels. A multiple-choice prediction is scored by the extracted option letters. If no letter can be extracted, an LLM judge (Qwen3-32B) reads the response; if it still fails to extract a valid option, the prediction is marked incorrect. For short-answer questions in the GeoOPD-ID150 task, exact matching is prioritized, with Qwen3-32B adjudication as a fallback.

Direct inference uses the student prompt from training. No evidence, reference answer, or crop instruction is added.

Inference prompt

System:

> You are a remote-sensing visual reasoning assistant.
> Write reasoning in `<think>...</think>` and the final answer in `<answer>...</answer>`.

User:

```
<image>
{question}
```

MLLM Output:

```
<think>...</think>
<answer>...</answer>
```

A.3 Data Construction

The construction procedure is described in the main text. This section records the resulting distributions and the prompts used to propose and filter questions.

A.3.1 GeoOPD-10K

GeoOPD-10K contains 10,000 questions: 4,000 short answers and 2,000 long answers synthesized from LRS-GRO images, and 4,000 multiple-choice questions retained from CHOICE. Table 5 groups them by evidence organization.

Table 5: Evidence organizations in GeoOPD-10K.

Evidence organization	Questions	Share
Single-object	3,278	32.8%
Paired-object	2,065	20.7%
Regional	1,702	17.0%
Instance-set	406	4.1%
Global	2,549	25.5%
Total	10,000	100%

A.3.2 GeoOPD-ID150

As shown in Figure 5, GeoOPD-ID150 contains 150 questions from images and items held out from GeoOPD-10K. They are not used for training or validation. Accuracy is computed on 125 questions: 75 multiple-choice questions and 50 short-answer questions. The remaining 25 long-answer questions are qualitative only.

Of the 150 questions, 75 are held-out CHOICE items and 75 are synthesized on the embargoed LRS-GRO images. The synthesized subset contains 34 full-image questions, 31 regional crops, 3 counting crops, and 7 yes/no crops. Its task types are object category, color, material, shape, and state (43), spatial relation (7), visual comparison (6), region semantics (6), global scene (6), and sparse counting (7). The CHOICE subset covers attribute recognition (9), attribute comparison (7), object presence (9), object counting (10), spatial relationship (10), scene classification (9), environmental assessment (9), situation inference (9), and disaster discrimination (3).

A.3.3 Synthesis and filtering prompts

The writer and critic share two images. Image 1 is clean. Image 2 is the same view with internal node identifiers drawn on detector proposals. Those identifiers are alignment tools and are removed before training. Each box below is one call.

Image-First writer exemplar

The writer drafts one question for one task type. It must skip the category when the image cannot support a unique answer.

Input:

Write ONE short remote-sensing QA from the images. English only. Image 1 is the clean scene. Image 2 only helps match boxes. Rex labels may be wrong. If this category is not visually clear, return {"skip": true}.

Category: sparse_count. Count every listed instance of one class. They are all fully inside. The answer is that integer.

Teacher evidence (bound, then cover): name the search area, state what counts as one instance, and check the boundary. Do not write the integer in the steps. Cite every listed node. Visual: mark all instances in one view; do not split them into separate crops.

Reasoning: 3 teacher-only steps, in the order grounding, observation, inference. Say where to look and what to compare, and stop before the verdict. Never write the answer word.



Q: What color is the roof of the moored ship?
A: Blue



Q: How many buildings with bright roofs are scattered among the forest clearings?
A: 6



Q: What is the dominant land cover and main facility type in this coastal scene?
A: Dense forest surrounds a coastal industrial complex featuring reactor domes and water tanks.

Figure 5: GeoOPD-ID150 samples.

Referring: one short handle of 2–5 words, such as “the harbor basin”. Never write “the object” or a Rex identifier in the question. At most 22 words. Do not begin with “In the image”.

Nodes:


```
[{"node_id":"rex:0007","label":"ship","bbox_q1000":[612,531,657,
  705]},
 {"node_id":"rex:0008","label":"ship","bbox_q1000":[628,705,761,
  800]},
 {"node_id":"rex:0011","label":"ship","bbox_q1000":[657,705,760,
  800]}]
```


Output:


```
{"question":"How many ships are in the harbor basin?",
 "answer":"3",
 "evidence_node_ids":["rex:0007","rex:0008","rex:0011"],
```

```
 "reasoning_steps":[
   "Restrict the search to the basin enclosed by the three piers.",
   "Treat a separate visible hull at a pier as one ship.",
   "Check the open water away from the piers for any additional
     hull."],
 "confidence":0.91}
```

Image-First critic exemplar

The critic scores the draft against the same two images. It does not rewrite the question or the answer.

Input:

> Act as a strict independent visual critic. Reject a draft when the answer is uncertain, the question is longer than about 22 words, a cited node is missing or unused, or a reasoning step states the verdict. For counting, accept only when every cited instance is fully inside the frame and the answer equals that set. Return one decision per candidate.

Candidate 0 is the writer output above.

Output:

```
{"decisions":[{
  "candidate_index":0,
  "accepted":true,
  "answer_supported":true,
  "question_clear":true,
  "evidence_supported":true,
  "privilege_supported":true,
  "count_coverage_complete":true,
  "reasons":["Three listed hulls are inside the basin; the steps do
    not state the count."]}]}
```

CHOICE evidence-writer exemplar

The official question and answer stay unchanged. The writer adds only teacher evidence. The official answer is private validation context and must not appear in the steps.

Input:

> Keep the official multiple-choice question unchanged. Write 2–4 teacher-only sentences. Never write option letters, option text, or the answer. Direction words may locate an object relative to an unrelated landmark, but must not state the relation asked by the question.
>
> Leaf: `object_color`. Describe the visible surface of the referred object and stop before naming the color.
>
> Question stem: What is the color of the cargo ship beside the quay?
>
> Options: `{"A":"red","B":"blue","C":"green","D":"white"}`
>
> Official answer: B. Official answer text: blue.
>
> Nodes: `[{"node_id":"rex:0004","label":"ship"}]`

Output:

```
{"steps":[
   "Locate the cargo ship moored along the stone quay.",
   "Inspect the hull surface above the waterline."],
 "evidence_node_ids":["rex:0004"]}
```

The CHOICE critic rejects a step that supports a different option, reverses a direction, leaks option text, or invents an object. Deterministic rules then remove student questions that contain the answer, unresolved references, duplicate questions, residual internal markers, and counting sets that do not cover every listed instance.

A.4 Inference Latency

Table 6 reports per-question latency on all 150 GeoOPD-ID150 questions. Each GPU serves one question at a time, and a warmup question on each GPU is discarded. The recorded time is the wait for that question. ZoomEarth's time adds the crop request when the model issues one.

Table 6: Latency on GeoOPD-ID150, in seconds. Tool calls are additional model requests beyond the initial forward pass.

Model	Mean	Median	Tool calls
GeoOPD-4B	0.64	0.58	0
Qwen2.5-VL-3B	0.66	0.55	0
Qwen3-VL-4B	0.82	0.75	0
ZoomEarth-3B	2.03	1.92	0.63
RSThinker	3.13	3.25	0
Qwen3.5-4B	12.10	4.38	0

GeoOPD-4B answers each question in one forward pass, with a mean latency of 0.64 seconds. This is close to Qwen2.5-VL-3B at 0.66 seconds and lower than its Qwen3-VL-4B backbone at 0.82 seconds. ZoomEarth makes 0.63 additional tool calls per question, and its mean latency is 2.03 seconds. RSThinker spends 3.13 seconds on its longer reasoning trace. Qwen3.5-4B has a mean of 12.10 seconds and a median of 4.38 seconds, because thinking mode produces much longer outputs. In this setup, GeoOPD achieves lower inference latency than the crop-based ZoomEarth baseline, consistent with the additional overhead introduced by crop requests.

A.5 Additional Case

Figure 6 compares GeoOPD with ZoomEarth on a position question. The yellow excavator lies on the shore along the bottom edge of the image. GeoOPD identifies the excavator beside the water at the bottom and selects option A.

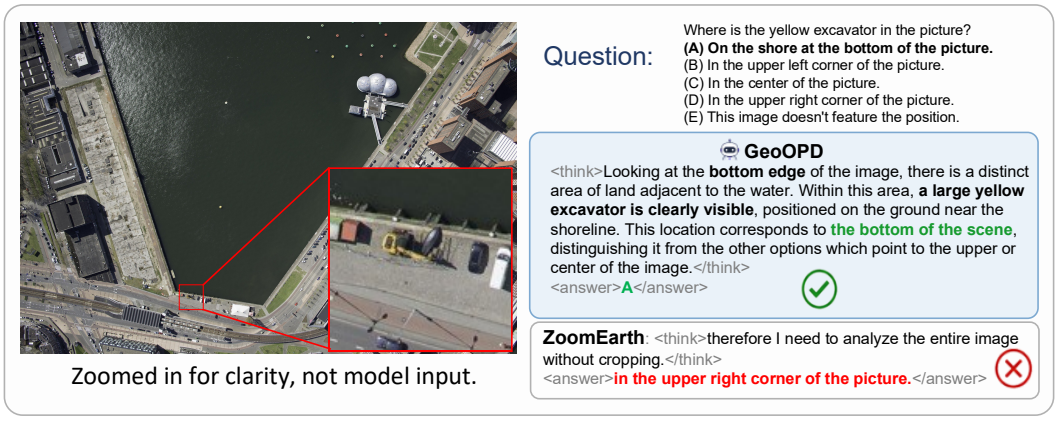


Figure 6: Case study on MME-RealWorld-RS.