GeoEPS: Training-Free Evidence Presentation and Selection for Ultra-High-Resolution Remote Sensing Understanding

Anonymous authors
Paper under double-blind review

Abstract

Ultra-high-resolution (UHR) remote-sensing (RS) imagery poses a fundamental challenge for vision-language models (VLMs): task-relevant evidence may occupy only a tiny fraction of a large scene, yet making that evidence more accessible does not ensure its effective use in the final answer. We identify this discrepancy as an evidence-to-decision gap, which is often observed in visual question answering (VQA) tasks for UHR-RS imagery. Through controlled comparisons, we find that how a referenced region is presented matters beyond simply locating it. To address this gap, we introduce GeoEPS, a training-free framework that coordinates evidence presentation and selection with frozen VLMs. When the input provides an explicit spatial cue, Evidence-aware Presentation (EAP) uses it as an evidence prior and combines local context, textual alignment, unfilled marking, and query-conditioned presentation. For inputs without an explicit spatial cue, Competitive Evidence Selection (CES) searches local evidence candidates and compares them under competing answer hypotheses. On XLRS-Bench and RSHR-Bench, our GeoEPS *substantially* improves Qwen3-VL-8B's average task accuracy from 48.8% to 61.0% and from 41.3% to 53.7%, respectively. Extensive experiments further demonstrate that these gains can extend to other frozen general-purpose and RS-specific VLMs without parameter updates or UHR-specific retraining.

1 Introduction

Vision-language models (VLMs) increasingly support remote-sensing (RS) image understanding and visual question answering (VQA) (Liu et al., 2024; Hu et al., 2025; Kuckreja et al., 2024; Zhang et al., 2024; Muhtar et al., 2025; Soni et al., 2025). Ultra-high-resolution (UHR) imagery preserves wide-area context and fine local detail, but the evidence needed to answer a question may occupy only a tiny region. XLRS-Bench images average approximately $8.5\text{k} \times 8.5\text{k}$ pixels, and RSHR-Bench evaluates scenes with long sides of at least 4k pixels (Wang et al., 2025b; Dang et al., 2025). VLMs encode images into visual tokens and operate within a limited visual-token budget. Fitting UHR images within this budget may require downsampling, which can obscure small objects and their attributes. The challenge is to provide VLMs with fine-grained, query-relevant evidence while preserving the scene context needed to interpret it.

Recent UHR-RS methods improve visual access through token reduction, retrieval, hierarchical search, and active perception (Wang et al., 2025a; Dang et al., 2026; Zhang et al., 2025; Zhou et al., 2025; Liu et al., 2025; Wang et al., 2026a;b). Yet deciding where to look is only part of the problem. For region-specific questions, the input can include an explicit spatial cue, such as bounding-box coordinates or a visual marker, that identifies the intended region among visually similar objects (Wang et al., 2025b; Kuckreja et al., 2024; Yang et al., 2023; Cai et al., 2024; You et al., 2024). For example, a user can select a region of interest and ask a VLM about its contents, a form of region-conditioned dialogue already supported by grounded RS VLMs (Kuckreja et al., 2024). Both the native baseline and GeoEPS receive these spatial cues as part of the original task input, and GeoEPS uses them as evidence priors. Such a prior specifies which region to interpret, but not *how its evidence should be presented*. Access to the region alone therefore does not ensure that the model can use its evidence effectively.

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.

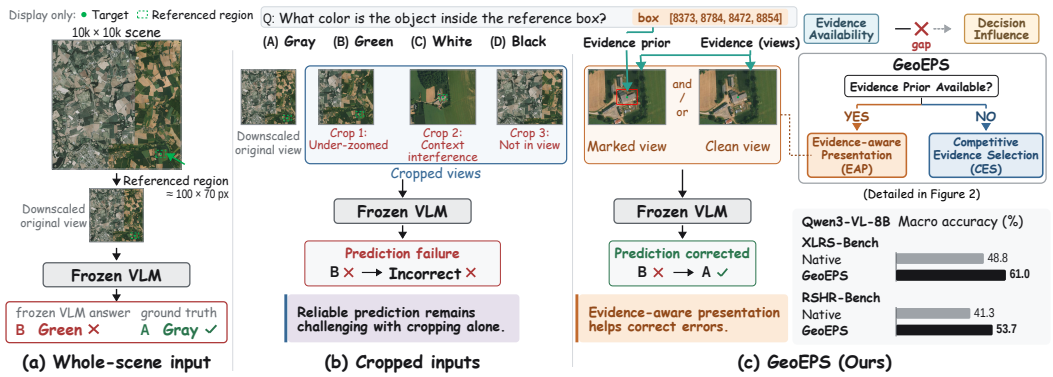


Figure 1: **Evidence-to-decision gap.** (a) A frozen VLM misidentifies the referenced object's color in the downsampled scene. (b) Crop failures show that cropping alone does not ensure a correct answer. (c) Given the supplied spatial cue, EAP presents the referenced region and corrects the prediction.

Fig. 1 illustrates this *evidence-to-decision gap*: making relevant evidence locally accessible does not necessarily make it reliably usable by the model. The crop examples in (b) expose distinct failure modes: the target may remain too small, become ambiguous among surrounding regions, or fall outside the selected view. These cases illustrate both evidence-access and presentation failures. EAP addresses the subsequent presentation problem by combining local context, an explicit reference, and query-aligned input construction. The example therefore motivates treating evidence access and evidence presentation as distinct design problems.

To address this evidence-to-decision gap, we propose GeoEPS, a training-free framework with two complementary routes depending on the availability of an explicit spatial cue. The Evidence-aware Presentation (EAP) module treats a supplied spatial cue as an evidence prior rather than merely a localization hint. The key idea is to transform available reference information into a query-conditioned visual presentation that helps a frozen VLM use the corresponding evidence effectively. For numeric pointers, EAP incorporates surrounding context, aligns the textual reference with the local view, and selects marked, clean, or joint presentations according to the query. For markers already present in the image, it preserves the marker and the original query while preserving the surrounding context.

When an explicit spatial cue is unavailable, the Competitive Evidence Selection (CES) module searches for relevant evidence candidates from the scene. It first performs sparse evidence search to construct a limited candidate pool and uses self-verification to assess whether each candidate provides sufficient evidence for answering. The highest-scoring candidate becomes the incumbent, and the highest-ranked distinct evidence candidate yielding a different answer hypothesis becomes the challenger. CES performs a pairwise comparison between the incumbent and challenger using frozen discriminative verification. Whereas self-verification assesses each view's sufficiency, discriminative verification asks which evidence better distinguishes the competing answer hypotheses. The selected candidate's answer becomes the final prediction.

Both modules operate on the frozen VLM without parameter updates or ground-truth answers. EAP addresses how available evidence is presented, whereas CES addresses which evidence to acquire and select. Experiments on XLRS-Bench and RSHR-Bench demonstrate consistent gains across diverse frozen general-purpose and RS-specific VLMs. GeoEPS improves Qwen3-VL-8B's average task accuracy from 48.8% to 61.0% on XLRS-Bench and from 41.3% to 53.7% on RSHR-Bench without parameter updates or UHR-specific retraining. Our contributions are summarized as:

- We identify the evidence-to-decision gap in UHR-RS VQA, distinguishing access to relevant evidence from its effective use in the final prediction.
- We introduce GeoEPS, a training-free framework for evidence presentation and selection. EAP transforms spatial cues into evidence priors for query-conditioned presentation, while CES performs competitive selection among evidence candidates.
- We demonstrate the effectiveness and portability of GeoEPS across frozen general-purpose and RS-specific VLMs on two UHR-RS benchmarks, achieving Qwen3-VL-8B gains of 12.2 and 12.4 percentage points without model-specific adaptation or UHR-specific retraining.

2 Related Work

High-resolution remote-sensing VLMs. Vision-language pretraining and instruction tuning support general-purpose visual understanding (Radford et al., 2021; Liu et al., 2023), while domain-specific models extend these capabilities to RS imagery and reasoning (Bazi et al., 2024; Pang et al., 2025; Li et al., 2026; Liu et al., 2026b). UHR imagery introduces a further challenge: small targets must remain distinguishable within wide-area context, while the number of visual tokens grows with input resolution. Tiling and multi-resolution encoding preserve local detail (Ye et al., 2023; Li et al., 2024; Guo et al., 2025; Luo et al., 2025; Chen et al., 2024; Wu et al., 2024), while compact visual representations reduce token cost (Huang et al., 2025; Li et al., 2025b). GeoLLaVA-8K prunes redundant background and retains object-centric tokens (Wang et al., 2025a); UHR-BAT uses query-guided multi-scale selection with region-wise preservation and merging (Dang et al., 2026). Although these designs reduce memory costs, aggressive compression can still remove fine detail, motivating selective inspection.

Visual search and active perception. To recover detail through additional observations, region-guided inference (Wu & Xie, 2024; Shao et al., 2024) and retrieval-based or active-perception methods inspect selected regions (Zhang et al., 2025; Liu et al., 2025). Unlike uniform tiling, these methods allocate observations according to the question. ZoomSearch provides training-free patch search using relevance and evidence-sufficiency scores (Zhou et al., 2025). GeoEyes learns when to zoom and stop (Wang et al., 2026a). GeoLens extends zooming with grounding and line-drawing tools (Wang et al., 2026b). Both use supervised fine-tuning and reinforcement learning for tool decisions. Selective inspection adds model or tool calls, while learned policies require training. GeoEPS requires no parameter updates: CES compares alternative evidence when views support different answers, while EAP bypasses search when a supplied spatial cue identifies the region.

Evidence organization and use. Even after acquisition, isolated crops can lose spatial relationships, and relevant tokens can be diluted during reasoning. WeaveEarth combines complementary crops with location metadata and a layout-preserving evidence board to reduce spatial fragmentation in global-local reasoning (Ma et al., 2026). Thinking-Once addresses dilution of encoded visual evidence by preserving question-relevant tokens and supporting context at intermediate layers (Mao et al., 2026). Both are training-free, but focus on cross-view organization or internal token routing. How a supplied spatial reference should shape VLM inputs remains a distinct design question.

Spatial reference and visual prompting. To disambiguate the target of a question, general-domain methods use visual marks or region representations. Set-of-Mark labels segmentation-derived regions (Yang et al., 2023), while learned interfaces interpret drawn prompts or combine coordinates with region features (Cai et al., 2024; You et al., 2024). These methods support region-level interaction, but identifying a referent does not determine the context or visual presentation suited to the question. A tight crop can remove informative context, whereas a wider crop can leave several objects as plausible referents. EAP addresses this context-reference trade-off by adapting the region's presentation to the question, without retraining the VLM or adding an auxiliary perception model.

3 Method

3.1 Overall Framework

Given visual input $\mathbf{I}$, a query q, and allowed answer options $\mathcal{Y}(q)$, GeoEPS constructs the evidence input for a frozen VLM F to answer the query. The input may contain one or more images, with or without an explicit spatial cue. As shown in Fig. 2, the EAP module adapts the presentation of a region identified by such a cue, whereas the CES module searches for and selects local evidence when no cue is supplied.

Let $r \in \{\mathrm{POINTER}, \mathrm{MARKER}, \varnothing\}$ denote the cue type: $\mathrm{POINTER}$ for a numeric pointer specified by a bounding box in the query, MARKER for an existing visual marker, and $\varnothing$ when neither is present. GeoEPS uses the following evidence-prior routing rule:

$$\rho(r) = \begin{cases} \mathrm{EAP}, & r \in \{\mathrm{POINTER}, \mathrm{MARKER}\}, \\ \mathrm{CES}, & r = \varnothing. \end{cases} \tag{1}$$

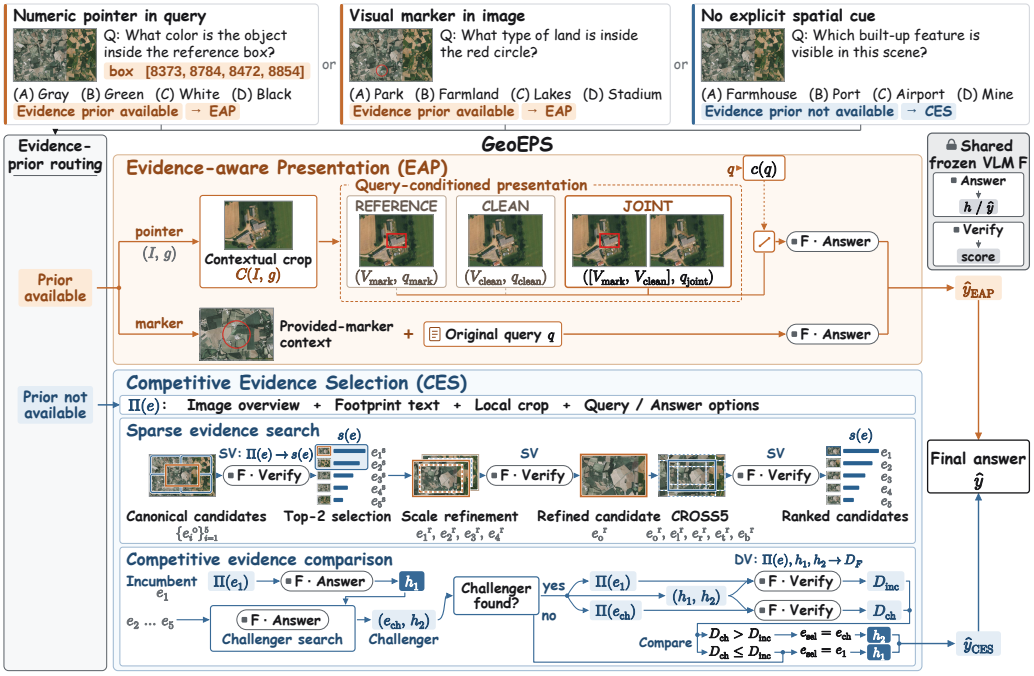


Figure 2: **Overall Framework of GeoEPS.** Zoom in for better visualization. Evidence-prior routing selects Evidence-aware Presentation (EAP) when spatial cues are available and Competitive Evidence Selection (CES) otherwise. EAP transforms spatial cues into localized evidence inputs through contextual cropping, reference alignment, visual marking, and query-conditioned selection, while CES searches and compares evidence candidates before selection. All calls use the frozen VLM.

This routing depends on the supplied cue, not the benchmark task label. Throughout, all model calls use F without parameter updates or ground-truth answers. We denote by $\mathrm{Ans}_F(X)$ the parsed answer generated by F from input X using the benchmark answering interface.

3.2 Evidence-Aware Presentation Module

A spatial cue identifies an intended region, but does not specify how much context to show or how to make the region clear within that context. The EAP module therefore coordinates contextual cropping, textual alignment, and query-conditioned presentation.

Contextual cropping and visual marking. A tight crop may remove informative context, whereas a wider crop can contain several plausible referents. For a numeric pointer represented by box geometry g in the coordinate frame stated in the query, $\mathcal{C}$ denotes the contextual cropping operator. It first maps the box coordinates to the coordinate system of the corresponding input image, then extracts an enlarged local window around the mapped box to include nearby visual context. We denote by $\widetilde{g}$ the same referenced box expressed in the coordinate system of the resulting contextual crop. The EAP module then constructs a clean view V_{clean} and its marked counterpart V_{mark}:

$$V_{\mathrm{clean}} = \mathcal{C}(\mathbf{I}, g), \qquad V_{\mathrm{mark}} = \mathcal{M}(V_{\mathrm{clean}}, \widetilde{g}), \tag{2}$$

where $\mathcal{M}$ draws an unfilled red outline immediately outside $\widetilde{g}$ before the resulting view is passed through the VLM's image processor. The outline identifies the intended region without covering its interior. The two views use the same contextual crop and therefore preserve the same surrounding context, differing only by the added outline.

Textual reference alignment and query-conditioned presentation. The textual reference should match the visual input presented to the model. We denote the three aligned query variants as q_{mark}, q_{clean}, and q_{joint}, corresponding to the marked view, clean view, and joint pair of views, respectively. For the marked view, q_{mark} replaces the numeric pointer with a reference to the region outlined by the red rectangle. The clean variant q_{clean} instead refers to the corresponding contextual crop. For the joint pair, q_{joint} uses the same red-rectangle reference as q_{mark} and additionally specifies that the first image is V_{mark} and the second is V_{clean}. Answer choices and their order are preserved.

We define a fixed query-conditioned presentation selector $c(\cdot)$ that maps the original query to one of three presentation modes, $c(q) \in \{\text{REFERENCE}, \text{CLEAN}, \text{JOINT}\}$. Let Φ_{EAP} denote the EAP presentation passed to F, including the selected view(s), aligned query, and answer choices:

$$\Phi_{\text{EAP}} = \begin{cases} (V_{\text{mark}}, q_{\text{mark}}, \mathcal{Y}(q)), & c(q) = \text{REFERENCE}, \\ (V_{\text{clean}}, q_{\text{clean}}, \mathcal{Y}(q)), & c(q) = \text{CLEAN}, \\ ([V_{\text{mark}}, V_{\text{clean}}], q_{\text{joint}}, \mathcal{Y}(q)), & c(q) = \text{JOINT}. \end{cases} \tag{3}$$

Within the pointer branch, queries matching the object-category pattern use REFERENCE, those matching the motion-state pattern use CLEAN, and those matching the color pattern use JOINT. If none or more than one of these patterns matches, $c(q)$ defaults to JOINT. The selector $c(\cdot)$ uses only the original query text q, not task IDs or answer correctness. Appendix B gives the exact matching rules and comparisons with fixed presentations.

Existing visual markers. For an existing visual marker, the EAP module extracts a marker-preserving contextual view V_{prov} that retains both the marker and its surrounding context. Because the referent is already marked, the original query q is preserved without pointer-specific rewriting:

$$\Phi_{\text{EAP}} = (V_{\text{prov}}, q, \mathcal{Y}(q)). \tag{4}$$

This branch neither draws an additional marker nor applies the query-conditioned selector $c(\cdot)$. For either EAP branch, the final answer is $\hat{y} = \text{Ans}_F(\Phi_{\text{EAP}})$.

3.3 Competitive Evidence Selection Module

Without a supplied spatial cue ($r = \varnothing$), the CES module searches for and ranks evidence candidates using a visual-evidence sufficiency score. It then compares the top-ranked evidence candidate with the highest-ranked distinct evidence candidate that yields a different answer hypothesis, asking which evidence better distinguishes the two hypotheses.

Evidence candidates and self-verification. For a source image $\mathbf{I}$, an evidence candidate e specifies a local crop $\mathbf{I}^{\text{loc}}(e)$ and its spatial footprint $f(e)$, which locates the local crop within the source image. We denote the model input constructed for evidence candidate e by

$$\Pi(e) = (\mathbf{I}, f(e), \mathbf{I}^{\text{loc}}(e), q, \mathcal{Y}(q)), \tag{5}$$

where $\mathbf{I}$ provides the image overview and $f(e)$ describes the local crop's bounding box as text in normalized source-image coordinates. To rank the evidence candidates, CES evaluates each $\Pi(e)$ with the frozen VLM F using self-verification (SV). Within this input, the image overview $\mathbf{I}$ and local crop $\mathbf{I}^{\text{loc}}(e)$ pass through the VLM's image processor, while $f(e)$, q, and $\mathcal{Y}(q)$ are supplied as text. The fixed template $\mathcal{T}_{\text{SV}}$ asks F to answer YES or NO according to whether $\Pi(e)$ contains enough visual evidence to answer q. We define the visual-evidence sufficiency score of e as:

$$s(e) = \log P_F(\text{YES} \mid \mathcal{T}_{\text{SV}}(\Pi(e))) - \log P_F(\text{NO} \mid \mathcal{T}_{\text{SV}}(\Pi(e))). \tag{6}$$

Here P_F denotes the conditional probability assigned by F to the specified verifier response. It is noted that a higher $s(e)$ indicates greater model-assessed sufficiency of the visual evidence for answering q, not a calibrated probability of answer correctness.

Sparse evidence search. The CES module begins with five overlapping canonical evidence candidates at the four corners and center of the source image, denoted by $\mathcal{E}^{\text{o}} = \{e_i^{\text{o}}\}_{i=1}^{5}$. It applies SV to each input $\Pi(e_i^{\text{o}})$ to obtain its visual-evidence sufficiency score, then sorts the evidence candidates as:

$$e_i^{\text{s}} = e_{\pi(i)}^{\text{o}}, \qquad s(e_1^{\text{s}}) \geq s(e_2^{\text{s}}) \geq \cdots \geq s(e_5^{\text{s}}),$$

where $\pi(i)$ maps rank i to the original index, and superscripts o and s denote the original and sorted evidence candidates, respectively. Among these, the two highest-scoring candidates, e_1^{s} and e_2^{s}, are refined at two scales. The scaling operator $\text{Scale}(e, \alpha)$ scales the width and height of the window defining the local crop $\mathbf{I}^{\text{loc}}(e)$ by a factor α while keeping its center fixed. We denote the four refined evidence candidates using superscript r:

$$e_1^{\text{r}} = \text{Scale}(e_1^{\text{s}}, 1.00), \quad e_2^{\text{r}} = \text{Scale}(e_1^{\text{s}}, 0.75), \quad e_3^{\text{r}} = \text{Scale}(e_2^{\text{s}}, 1.00), \quad e_4^{\text{r}} = \text{Scale}(e_2^{\text{s}}, 0.75).$$

The refined evidence candidates are then compared by their SV scores, and the highest-scoring one is selected as the anchor for subsequent translation refinement:

$$e_{\text{o}}^{\text{r}} \in \underset{e \in \{e_1^{\text{r}}, \ldots, e_4^{\text{r}}\}}{\arg\max}\ s(e).$$

The CES module then explores a five-position neighborhood consisting of the unshifted anchor $e_{\mathrm{o}}^{\mathrm{r}}$ and four evidence candidates obtained by translating its crop window:

$$\mathrm{CROSS5}(e_{\mathrm{o}}^{\mathrm{r}}) = \{e_{\mathrm{o}}^{\mathrm{r}}, e_{\mathrm{l}}^{\mathrm{r}}, e_{\mathrm{r}}^{\mathrm{r}}, e_{\mathrm{t}}^{\mathrm{r}}, e_{\mathrm{b}}^{\mathrm{r}}\}, \tag{7}$$

where the subscript o denotes the unshifted origin of the translation refinement, and $\mathrm{l}, \mathrm{r}, \mathrm{t}, \mathrm{b}$ denote the left, right, top, and bottom shifts, respectively. Horizontal shifts are 20% of the local crop width, and vertical shifts are 20% of its height. At image boundaries, translated crop positions are clamped to the valid image range while preserving the window size.

After SV scoring, the five CROSS5 evidence candidates are reordered as $e_1, \ldots, e_5$ such that

$$s(e_1) \geq s(e_2) \geq \cdots \geq s(e_5).$$

The evidence candidate with the highest SV score, e_1, becomes the *incumbent*, the provisional evidence choice carried into competitive selection.

Challenger search. The CES module first obtains the incumbent hypothesis $h_1 = \mathrm{Ans}_F(\Pi(e_1))$. It then scans $e_2, \ldots, e_5$ in descending order of $s(e)$ until it finds the first eligible evidence candidate supporting an answer hypothesis different from h_1. During this scan, boundary clamping can cause different CROSS5 shifts to yield the same realized local crop as e_1. Let $\mathrm{iid}(e)$ denote the realized evidence identity used to exclude such duplicates. For each remaining evidence candidate e in turn, the module obtains $h_2 = \mathrm{Ans}_F(\Pi(e))$. The first evidence candidate satisfying

$$\mathrm{iid}(e) \neq \mathrm{iid}(e_1), \qquad h_2 \neq h_1$$

becomes the *challenger* e_{ch}. Here hypotheses are parsed answers in $\mathcal{Y}(q)$. If no challenger is found, we set the selected evidence $e_{\mathrm{sel}} = e_1$ and the final prediction $\hat{y} = h_1$. Thus, e_{ch} is the highest-ranked distinct evidence candidate whose hypothesis disagrees with h_1, rather than necessarily e_2.

Discriminative verification and final prediction. While SV assesses whether an evidence candidate is sufficient to answer q, discriminative verification (DV) asks whether its evidence distinguishes the competing hypotheses h_1 and h_2. With the fixed template $\mathcal{T}_{\mathrm{DV}}$, F answers YES or NO to this discriminative question. We define the discriminative evidence score as:

$$D_F(\Pi(e); h_1, h_2) = \log P_F(\mathrm{YES} \mid \mathcal{T}_{\mathrm{DV}}(\Pi(e); h_1, h_2)) - \log P_F(\mathrm{NO} \mid \mathcal{T}_{\mathrm{DV}}(\Pi(e); h_1, h_2)). \tag{8}$$

The incumbent and challenger are evaluated separately using the same ordered hypothesis pair:

$$D_{\mathrm{inc}} = D_F(\Pi(e_1); h_1, h_2), \qquad D_{\mathrm{ch}} = D_F(\Pi(e_{\mathrm{ch}}); h_1, h_2). \tag{9}$$

When a challenger exists, the selected evidence e_{sel} and final prediction $\hat{y}$ are jointly determined as:

$$(e_{\mathrm{sel}}, \hat{y}) = \begin{cases} (e_{\mathrm{ch}}, h_2), & D_{\mathrm{ch}} > D_{\mathrm{inc}}, \\ (e_1, h_1), & D_{\mathrm{ch}} \leq D_{\mathrm{inc}}. \end{cases} \tag{10}$$

4 Experiments

4.1 Experimental Setup

Datasets and metrics. We evaluate GeoEPS on two UHR remote-sensing VQA benchmarks: XLRS-Bench (Wang et al., 2025b), with images averaging approximately $8.5\mathrm{k} \times 8.5\mathrm{k}$ pixels, and a fixed single-turn subset of RSHR-Bench (Dang et al., 2025), with image long sides of at least 4k pixels. Both benchmarks cover perception and reasoning tasks. We report two aggregate accuracy metrics: Macro, the unweighted mean of task accuracies, and Micro, the sample-weighted mean of task accuracies. Macro_k averages the specified k tasks in a given analysis.

Models and inference. We evaluate six open-source general-purpose and six remote sensing frozen MLLMs. Literature-reported closed-source results ($^{\mathrm{L}}$) are included only for reference. We compare Native with GeoEPS, which uses evidence-prior routing: EAP for explicit spatial cues and CES for no-cue inputs. Numeric pointers are provided in XLRS OCC/OCL/OMS tasks and RSHR COL/SHP/DET/OC/RCN tasks, while XLRS RC/RLUC/RCCD tasks also contain visual markers. Native and GeoEPS use the same original inputs, including supplied spatial cues, while EAP changes how referenced evidence is presented. For ZoomEarth, GeoLens, and GeoEyes (*), which have native evidence acquisition, we apply only EAP. GeoLens† uses the released checkpoint with a reconstructed evaluation pipeline based on its public code and method description. All inference experiments use two NVIDIA B200 GPUs. Further details are provided in Appendix A.

Table 1: **Experimental results on XLRS-Bench.** Each evaluated model is reported with Native and shaded GeoEPS inference. Δ denotes the gain over Native. Literature-reported results are marked with ^{L}. For models marked with *, EAP is combined with native evidence acquisition instead of CES. GeoLens† uses the released checkpoint with our reconstructed evaluation pipeline.

Model	Inference	Perception								Reasoning					Overall			
		OC	RC	OLUC	RLUC	OCC	OCL	OMS	OSR	AD	ECR	RP	RCCD	CCR	Macro	Δ	Micro	Δ
Closed-source MLLMs																		
Claude 3.7 Sonnet^{L}		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	–	35.1	–
GPT-5.2^{L}		30.0	37.0	17.0	70.5	43.0	41.4	68.3	34.0	74.0	76.0	52.0	36.7	38.0	47.5	–	44.2	–
Open-source MLLMs																		
LLaVA-OV-7B	Native	20.0	38.0	2.0	71.0	36.3	33.9	65.0	25.8	75.0	83.0	24.0	48.3	34.0	42.8	–	37.9	–
(Li et al., 2025a)	**GeoEPS**	28.3	51.0	2.0	86.0	87.8	89.5	83.3	26.4	77.0	80.0	23.0	50.0	39.0	**55.6**	**+12.8**	**67.9**	**+30.0**
Idefics3-8B	Native	30.0	37.0	2.0	67.5	36.5	33.4	65.0	30.2	67.0	80.0	31.0	31.7	36.0	42.1	–	38.1	–
(Laurençon et al., 2024)	**GeoEPS**	36.7	46.0	2.0	83.0	81.4	79.0	71.7	28.4	66.0	81.0	33.0	46.7	32.0	**52.8**	**+10.7**	**63.1**	**+25.0**
InternVL3-8B	Native	36.7	35.0	2.0	73.5	43.2	31.0	63.3	23.6	75.0	79.0	34.0	36.7	38.0	43.9	–	39.1	–
(Zhu et al., 2025)	**GeoEPS**	28.3	48.0	7.0	82.5	90.1	88.5	85.0	26.6	77.0	75.0	32.0	31.7	49.0	**55.4**	**+11.5**	**68.2**	**+29.1**
Ministral 3 14B	Native	26.7	34.0	17.0	69.0	36.6	33.1	65.0	35.0	74.0	81.0	31.0	28.3	45.0	44.3	–	39.8	–
(Liu et al., 2026a)	**GeoEPS**	26.7	43.0	27.0	76.0	69.1	69.3	71.7	33.2	72.0	83.0	30.0	23.3	44.0	**51.4**	**+7.1**	**58.3**	**+18.5**
MiniCPM-V 4.5	Native	25.0	42.0	35.0	77.5	39.6	32.1	65.0	25.0	78.0	82.0	24.0	35.0	47.0	46.7	–	40.2	–
(Yu et al., 2026)	**GeoEPS**	26.7	53.0	33.0	84.0	91.8	86.3	85.0	24.6	76.0	82.0	29.0	26.7	50.0	**57.5**	**+10.8**	**68.9**	**+28.7**
Qwen3-VL-8B	Native	16.7	43.0	23.0	77.5	46.6	41.1	66.7	30.6	73.0	82.0	41.0	41.7	51.0	48.8	–	45.4	–
(Bai et al., 2025)	**GeoEPS**	28.3	59.0	29.0	85.5	92.5	88.0	90.0	32.4	71.0	82.0	39.0	41.7	54.0	**61.0**	**+12.2**	**71.7**	**+26.3**
Remote Sensing MLLMs																		
GeoChat	Native	16.7	29.0	1.0	23.0	21.1	16.8	35.0	23.8	36.0	48.0	46.0	–	21.0	26.4	–	22.5	–
(Kuckreja et al., 2024)	**GeoEPS**	28.3	29.0	2.0	26.0	21.4	16.8	35.0	32.4	58.0	66.0	46.0	–	44.0	**33.7**	**+7.3**	**26.6**	**+4.1**
EarthDial	Native	21.7	45.0	2.0	35.5	32.8	32.0	58.3	25.6	65.0	71.0	34.0	40.0	44.0	39.0	–	34.1	–
(Soni et al., 2025)	**GeoEPS**	18.3	43.0	2.0	70.5	75.4	76.9	78.3	25.0	60.0	70.0	25.0	43.3	33.0	**47.8**	**+8.8**	**58.5**	**+24.4**
RSCoVLM-7B	Native	25.0	30.0	2.0	61.5	35.5	44.9	65.0	29.6	70.0	77.0	35.0	46.7	49.0	43.9	–	40.9	–
(Li et al., 2026)	**GeoEPS**	33.3	57.0	2.0	80.0	88.8	84.0	78.3	30.2	72.0	76.0	31.0	45.0	48.0	**55.8**	**+11.9**	**67.3**	**+26.4**
ZoomEarth	Native	38.3	36.0	2.0	60.0	32.9	30.1	60.0	24.6	66.0	73.0	46.0	33.3	49.0	42.4	–	35.6	–
(Liu et al., 2025)	**GeoEPS***	38.3	45.0	2.0	69.5	69.3	64.9	86.7	24.6	66.0	73.0	46.0	26.7	49.0	**50.8**	**+8.4**	**55.4**	**+19.8**
GeoLens†	Native	28.3	39.0	45.0	78.0	49.8	44.4	80.0	37.2	68.0	77.0	56.0	51.7	55.0	54.6	–	49.7	–
(Wang et al., 2026b)	**GeoEPS***	28.3	53.0	45.0	83.0	82.0	77.5	71.7	37.2	68.0	77.0	56.0	48.3	55.0	**60.2**	**+5.6**	**67.2**	**+17.5**
GeoEyes	Native	16.7	40.0	2.0	77.5	58.3	68.6	65.0	33.2	74.0	80.0	55.0	30.0	43.0	49.5	–	55.1	–
(Wang et al., 2026a)	**GeoEPS***	16.7	55.0	2.0	77.5	79.6	74.8	78.3	33.2	74.0	80.0	55.0	41.7	43.0	**54.7**	**+5.2**	**63.2**	**+8.1**

4.2 Main Results

Results on XLRS-Bench. Table 1 shows that GeoEPS improves both Macro and Micro for all twelve evaluated models. Macro gains range from 7.1 to 12.8 points for open-source general-purpose MLLMs and from 5.2 to 11.9 points for remote sensing MLLMs. In particular, Qwen3-VL-8B reaches **61.0%** Macro and **71.7%** Micro, the highest overall scores among these models, compared with 48.8% and 45.4% for Native. For numeric-pointer questions in OCC/OCL/OMS, Qwen3's accuracies increase from 46.6/41.1/66.7% to 92.5/88.0/90.0%, gains of 45.9/46.9/23.3 points. Since Native and GeoEPS receive the same original spatial cues, these improvements show the benefit of EAP's full evidence-aware construction, which uses the supplied spatial cue as an evidence prior for contextual cropping, textual reference alignment, and query-conditioned visual presentation. The same strong gains extend to remote sensing MLLMs. RSCoVLM's OCC/OCL/OMS accuracies increase from 35.5/44.9/65.0% to 88.8/84.0/78.3%, raising Macro from 43.9% to 55.8%, and EarthDial's OCC/OCL/OMS improve from 32.8/32.0/58.3% to 75.4/76.9/78.3%, raising Macro from 39.0% to 47.8%. In contrast, GeoChat shows a different pattern. Its numeric-pointer OCC/OCL/OMS tasks change little with EAP, whereas CES substantially improves the no-cue OSR/AD/ECR/CCR tasks from 23.8/36.0/48.0/21.0% to 32.4/58.0/66.0/44.0%. This suggests that GeoChat's larger recoverable bottleneck lies in evidence acquisition on no-cue questions, which CES partially alleviates through explicit evidence search followed by competitive evidence selection. Finally, ZoomEarth, GeoLens, and GeoEyes already include their own evidence-acquisition mechanisms, yet adding EAP further improves Macro by 8.4/5.6/5.2 points, respectively.

Results on RSHR-Bench. Table 2 shows that GeoEPS improves both Macro and Micro for all twelve evaluated models on RSHR-Bench. Open-source general-purpose MLLMs gain 8.8–12.4 Macro points, and remote sensing MLLMs gain 4.0–11.7 points. Qwen3-VL-8B achieves **53.7%** Macro and **54.0%** Micro, improving over Native by 12.4 and 16.1 points, respectively. Among remote sensing models, RSCoVLM-7B reaches the highest Macro of 50.5%, up from 38.8%. For numeric-pointer questions in COL/SHP/OC, Qwen3's accuracies increase from 57.0/21.0/30.0% to 92.0/79.0/83.5%. The gain is not specific to Qwen3, with shape recognition (SHP) improving from 28.0% to 82.0%

Table 2: **Experimental results on the RSHR-Bench.** Each evaluated model is reported with Native and shaded GeoEPS inference. Δ denotes the gain over Native. Literature-reported results are marked with ^{L}. For models marked with *, EAP is combined with native evidence acquisition instead of CES. GeoLens† uses the released checkpoint with our reconstructed evaluation pipeline.

Model	Inference	Perception									Reasoning					Overall			
		COL	SHP	DET	OC	REL	OGD	RG	OCN	RCN	AD	FP	MRJC	MRJCS	OSJ	Macro	Δ	Micro	Δ
Closed-source MLLMs																			
GPT-5^{L}		29.0	10.0	23.0	23.0	37.0	24.5	31.4	20.0	23.0	74.0	58.0	35.0	34.0	66.0	34.8	–	29.4	–
GPT-4o^{L}		49.5	23.0	15.0	35.5	30.5	28.0	27.1	22.5	41.0	68.0	56.0	30.5	32.0	64.0	37.3	–	34.3	–
Open-source MLLMs																			
LLaVA-OV-7B	Native	50.0	27.0	27.0	30.5	45.0	26.0	30.0	17.0	29.0	68.0	32.0	50.0	36.0	64.0	38.0	–	34.7	–
(Li et al., 2025a)	**GeoEPS**	86.0	61.0	19.0	84.0	43.5	29.5	32.9	18.0	53.0	58.0	32.0	50.0	34.0	66.0	**47.6**	**+9.6**	**49.2**	**+14.5**
Idefics3-8B	Native	49.0	24.0	38.0	36.5	31.0	25.5	21.4	22.5	19.0	68.0	38.0	30.0	30.0	50.0	34.5	–	33.0	–
(Laurençon et al., 2024)	**GeoEPS**	84.5	47.0	32.0	76.0	29.0	25.0	30.0	19.5	37.0	68.0	38.0	30.0	36.0	54.0	**43.3**	**+8.8**	**44.6**	**+11.6**
InternVL3-8B	Native	47.0	28.0	24.0	47.5	42.0	34.5	21.4	27.5	40.0	64.0	40.0	65.0	40.0	46.0	40.5	–	38.5	–
(Zhu et al., 2025)	**GeoEPS**	89.0	82.0	25.0	83.0	42.5	27.5	28.6	30.5	48.0	68.0	40.0	65.0	34.0	54.0	**51.2**	**+10.7**	**52.3**	**+13.8**
Ministral 3 14B	Native	50.5	26.0	34.0	27.5	31.5	26.5	31.4	20.5	28.0	56.0	46.0	55.0	44.0	52.0	37.8	–	33.5	–
(Liu et al., 2026a)	**GeoEPS**	86.5	62.0	32.0	69.5	37.0	25.5	28.6	23.5	46.0	58.0	46.0	55.0	38.0	58.0	**47.5**	**+9.7**	**47.5**	**+14.0**
MiniCPM-V 4.5	Native	48.0	21.0	24.0	40.0	47.0	33.5	25.7	22.0	34.0	68.0	54.0	40.0	38.0	56.0	39.4	–	37.4	–
(Yu et al., 2026)	**GeoEPS**	92.5	75.0	25.0	79.0	47.0	30.5	22.9	22.5	50.0	66.0	54.0	40.0	34.0	64.0	**50.2**	**+10.8**	**51.9**	**+14.5**
Qwen3-VL-8B	Native	57.0	21.0	19.0	30.0	47.5	39.5	42.9	14.0	38.0	68.0	38.0	55.0	40.0	68.0	41.3	–	37.9	–
(Bai et al., 2025)	**GeoEPS**	92.0	79.0	24.0	83.5	42.0	39.0	35.7	18.0	58.0	70.0	52.0	55.0	36.0	68.0	**53.7**	**+12.4**	**54.0**	**+16.1**
Remote Sensing MLLMs																			
GeoChat	Native	44.5	11.0	28.0	36.0	21.5	24.5	21.4	19.0	28.0	46.0	34.0	–	32.0	30.0	28.9	–	28.3	–
(Kuckreja et al., 2024)	**GeoEPS**	56.0	34.0	33.0	59.5	26.0	27.0	22.9	22.5	32.0	56.0	34.0	–	38.0	44.0	**37.3**	**+8.4**	**37.1**	**+8.8**
EarthDial	Native	36.5	23.0	19.0	36.0	23.0	32.5	28.6	20.0	38.0	48.0	38.0	35.0	36.0	54.0	33.4	–	30.9	–
(Soni et al., 2025)	**GeoEPS**	86.0	48.0	21.0	52.0	21.5	31.5	24.3	17.5	39.0	56.0	38.0	35.0	34.0	62.0	**40.4**	**+7.0**	**40.5**	**+9.6**
RSCoVLM-7B	Native	49.5	22.0	13.0	33.0	40.5	33.5	55.7	20.5	27.0	70.0	46.0	50.0	26.0	56.0	38.8	–	35.5	–
(Li et al., 2026)	**GeoEPS**	84.0	68.0	23.0	79.0	44.5	33.5	44.3	22.0	45.0	74.0	46.0	50.0	30.0	64.0	**50.5**	**+11.7**	**50.9**	**+15.4**
ZoomEarth	Native	53.0	22.0	30.0	32.0	29.0	23.0	20.0	26.0	40.0	62.0	36.0	50.0	38.0	52.0	36.6	–	33.7	–
(Liu et al., 2025)	**GeoEPS***	82.0	55.0	32.0	57.0	29.0	23.0	20.0	26.0	42.0	62.0	36.0	50.0	38.0	52.0	**43.1**	**+6.5**	**42.8**	**+9.1**
GeoLens†	Native	57.0	16.0	22.0	33.5	48.0	38.0	32.9	16.5	41.0	60.0	38.0	50.0	36.0	70.0	39.9	–	37.7	–
(Wang et al., 2026b)	**GeoEPS***	89.5	64.0	28.0	70.0	48.0	38.0	32.9	16.5	56.0	60.0	38.0	50.0	36.0	70.0	**49.8**	**+9.9**	**50.8**	**+13.1**
GeoEyes	Native	77.5	56.0	27.0	38.0	48.5	32.0	41.4	22.5	48.0	64.0	36.0	40.0	38.0	64.0	45.2	–	44.4	–
(Wang et al., 2026a)	**GeoEPS***	82.5	61.0	36.0	67.0	48.5	32.0	41.4	22.0	56.0	64.0	36.0	40.0	38.0	64.0	**49.2**	**+4.0**	**50.0**	**+5.6**

for InternVL3 and from 22.0% to 68.0% for RSCoVLM. Moreover, RCN, another numeric-pointer task, improves from 38.0% to 58.0% for Qwen3 and increases across all twelve evaluated models, showing that the benefit can extend beyond local attribute recognition to regional counting. Together with the XLRS results, these findings show consistent EAP gains across multiple models and both benchmarks when explicit spatial cues are available. EAP also remains beneficial for ZoomEarth, GeoLens, and GeoEyes, which already perform native evidence acquisition, improving Macro by 6.5/9.9/4.0 points, respectively. For no-cue inputs, these models continue to rely on their own native evidence-acquisition mechanisms.

Table 3: **Ablation results on EAP/CES.**

Models	Native	EAP	GeoEPS (EAP+CES)
Open-source MLLMs			
Qwen3-VL-8B	48.8	59.6	61.0
LLaVA-OV-7B	42.8	54.7	55.6
InternVL3-8B	43.9	54.9	55.4
Remote Sensing MLLMs			
GeoChat	26.4	26.8	33.7
RSCoVLM-7B	43.9	55.3	55.8
GeoEyes*	49.5	54.7	–

Table 4: **Ablation results on EAP construction.**

ID	Construction	Macro_3	Δ_{step}	Δ_{cum}
E0	Native	51.5	–	–
E1	+ Tight crop	69.4	+17.9	+17.9
E2-R	+ Remapped numeric reference	74.1	+4.7	+22.6
E2	+ Textual reference alignment	75.8	+6.4	+24.3
E3	+ Contextual crop	76.8	+1.0	+25.3
E4-C	+ Centered marker outline	83.4	+6.6	+31.9
E4	+ Outside marker outline	85.8	+9.0	+34.3
E5	+ Query-conditioned presentation	**90.2**	+4.4	**+38.7**

4.3 Ablation Studies

We conduct ablations on XLRS-Bench with frozen Qwen3-VL-8B, except for Table 3.

Effect of the EAP and CES modules. Table 3 compares Native, EAP-only, and full GeoEPS. In the EAP-only setting, no-cue inputs use Native inference. Qwen3 improves from 48.8% to 59.6% Macro with EAP and to 61.0% with both modules. LLaVA-OV and InternVL3 show a similar pattern, with EAP providing most of the improvement and CES adding further gains. In contrast, GeoChat changes only from 26.4% to 26.8% with EAP but reaches 33.7% when CES is also enabled, indicating that

Table 5: **Ablation results on CES stages.**

ID	Operation	Evidence	Macro_9	Δ_{step}	Δ_{cum}
C0	Native	–	51.5	–	–
C1	Canonical search	e_1^{s}	51.8	+0.3	+0.3
C2	Scale refinement	e_{o}^{r}	52.3	+0.5	+0.8
C3	CROSS5 translation	e_1	52.9	+0.6	+1.4
C4	Competitive selection	e_{sel}	**53.4**	**+0.5**	**+1.9**

Table 6: **Numeric-pointer sensitivity.**

Translation	GeoEPS	Δ_{clean}	Scale	GeoEPS	Δ_{clean}
Clean	90.2	–	Clean	90.2	–
10%	88.8	-1.4	$1.1\times$	88.1	-2.1
20%	88.6	-1.6	$1.2\times$	88.8	-1.4
30%	87.5	-2.7	$1.5\times$	88.1	-2.1
50%	84.4	-5.8	$2.0\times$	85.3	-4.9

GeoChat benefits primarily from CES on no-cue inputs. Therefore, the relative contributions of EAP and CES can vary across models.

Effect of EAP construction choices. Table 4 traces the main EAP construction from using the supplied spatial cue as an evidence prior for tight cropping (E1), through textual reference alignment (E2), contextual cropping (E3), visual marking with an outside outline (E4), and query-conditioned presentation (E5). E2-R and E4-C serve as minimal-change control variants that adhere more directly to the supplied reference. E2-R uses the original numeric reference format and remaps only its coordinates to the crop frame. E4-C draws the marker directly along the boundary of the referenced box. In the main construction, E2 instead replaces the numeric box reference with text that directly refers to the local crop shown to the VLM, reducing the mismatch between the textual reference and the presented visual input. E4 places the outline fully outside the referenced box, preserving visual evidence within the referenced region, particularly near its boundary, while making the target location explicit. Starting from 51.5% Macro_3 at E0, tight cropping alone (E1) reaches 69.4%, a gain of 17.9 points, whereas the complete EAP (E5) reaches 90.2%, a cumulative gain of 38.7 points. Thus, tight cropping accounts for less than half of the total improvement, while the subsequent EAP construction choices collectively add another 20.8 points beyond E1. These paired comparisons further clarify the contribution of each design choice. Textual reference alignment reaches 75.8% at E2, compared with 74.1% for the coordinate-remapped numeric variant E2-R. Adding surrounding context gives 76.8% at E3. Visual marking with the outside outline reaches 85.8% at E4, compared with 83.4% for the boundary-centered variant E4-C. Finally, query-conditioned presentation further increases Macro_3 from 85.8% at E4 to 90.2% at E5, adding 4.4 points beyond the fixed marked presentation.

Effect of CES search and evidence comparison. Table 5 compares the successive CES stages. C0 is the no-CES baseline, using Native inference for no-cue inputs. Macro_9 is the unweighted mean of the full-task accuracies on AD, CCR, ECR, OC, OLUC, OSR, RC, RLUC, and RP. Starting from 51.5% at C0, it increases to 51.8% with canonical search (C1), 52.3% with scale refinement (C2), 52.9% with CROSS5 translation (C3), and 53.4% with competitive selection (C4), yielding a cumulative gain of 1.9 points. In particular, C3 and C4 share the same candidate pool and SV ranking, so the additional 0.5-point gain isolates the contribution of competitive evidence comparison.

Robustness to numeric-pointer perturbations. Table 6 evaluates GeoEPS under translation and scale perturbations of the supplied numeric pointers. Δ_{clean} denotes the change from the clean Macro_3 of 90.2%. GeoEPS remains close to its clean performance under moderate reference perturbations: 10% and 20% translations yield 88.8% and 88.6%, while scales from $1.1\times$ to $1.5\times$ remain between 88.1% and 88.8%. Even under substantially larger perturbations, performance remains 84.4% at 50% translation and 85.3% at $2.0\times$ scale, corresponding to decreases of only 5.8 and 4.9 points from clean. These results show that GeoEPS remains robust to considerable numeric-reference imprecision.

5 Conclusion

We presented GeoEPS, a training-free framework that addresses the evidence-to-decision gap in UHR-RS VQA with frozen VLMs. When an explicit spatial cue is available, EAP uses it as an evidence prior and coordinates contextual cropping, textual reference alignment, visual marking, and query-conditioned presentation. For no-cue inputs, CES searches local evidence candidates and compares an incumbent with a challenger before selection. Experiments on XLRS-Bench and RSHR-Bench show consistent gains across general-purpose and remote sensing MLLMs without parameter updates. Ablations further show that EAP gains extend well beyond tight cropping, and that both evidence search and competitive evidence comparison contribute to CES. Together, these results support evidence presentation and selection as complementary training-free mechanisms for making high-resolution visual evidence more usable by frozen VLMs.

AI USE STATEMENT

In this work, we used generative AI tools for aiding or polishing writing. We have not used generative AI tools for retrieval and discovery, research ideation or execution, drafting sections of the paper, generating synthetic datasets, or proving mathematical claims, and the remaining required disclosure tasks are not applicable to this work. We have reviewed all AI-assisted work. AI-assisted language suggestions were reviewed, revised, and incorporated by the authors where appropriate. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

Section 3 describes the GeoEPS inference procedure. Appendices A–C detail the evaluation protocols, input construction, search procedures, and verification settings. Appendices D–E describe the ablation settings and spatial-cue perturbation protocol. Inference configurations and evaluation manifests are recorded to support reproducibility. We plan to release inference and evaluation code, configurations, and evaluation manifests, subject to the licenses of the underlying models and datasets.

REFERENCES

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, Mei Li, Kaixin Li, Zicheng Lin, Junyang Lin, Xuejing Liu, Jiawei Liu, Chenglong Liu, Yang Liu, Dayiheng Liu, Shixuan Liu, Dunjie Lu, Ruilin Luo, Chenxu Lv, Rui Men, Lingchen Meng, Xuancheng Ren, Xingzhang Ren, Sibo Song, Yuchong Sun, Jun Tang, Jianhong Tu, Jianqiang Wan, Peng Wang, Pengfei Wang, Qiuyue Wang, Yuxuan Wang, Tianbao Xie, Yiheng Xu, Haiyang Xu, Jin Xu, Zhibo Yang, Mingkun Yang, Jianxin Yang, An Yang, Bowen Yu, Fei Zhang, Hang Zhang, Xi Zhang, Bo Zheng, Humen Zhong, Jingren Zhou, Fan Zhou, Jing Zhou, Yuanzhi Zhu, and Ke Zhu. Qwen3-VL Technical Report. *arXiv preprint arXiv:2511.21631v2*, 2025. URL https://arxiv.org/abs/2511.21631v2.

Yakoub Bazi, Laila Bashmal, Mohamad Mahmoud Al Rahhal, Riccardo Ricci, and Farid Melgani. RS-LLaVA: A Large Vision-Language Model for Joint Captioning and Question Answering in Remote Sensing Imagery. *Remote Sensing*, 16(9):1477, 2024. doi: 10.3390/rs16091477. URL https://doi.org/10.3390/rs16091477.

Mu Cai, Haotian Liu, Siva Karthik Mustikovela, Gregory P. Meyer, Yuning Chai, Dennis Park, and Yong Jae Lee. ViP-LLaVA: Making Large Multimodal Models Understand Arbitrary Visual Prompts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12914–12923, 2024. URL https://openaccess.thecvf.com/content/CVPR2024/html/Cai_ViP-LLaVA_Making_Large_Multimodal_Models_Understand_Arbitrary_Visual_Prompts_CVPR_2024_paper.html.

Zhe Chen, Weiyun Wang, Hao Tian, Shenglong Ye, Zhangwei Gao, Erfei Cui, Wenwen Tong, Kongzhi Hu, Jiapeng Luo, Zheng Ma, Ji Ma, Jiaqi Wang, Xiaoyi Dong, Hang Yan, Hewei Guo, Conghui He, Botian Shi, Zhenjiang Jin, Chao Xu, Bin Wang, Xingjian Wei, Wei Li, Wenjian Zhang, Bo Zhang, Pinlong Cai, Licheng Wen, Xiangchao Yan, Min Dou, Lewei Lu, Xizhou Zhu, Tong Lu, Dahua Lin, Yu Qiao, Jifeng Dai, and Wenhai Wang. How far are we to GPT-4V? Closing the gap to commercial multimodal models with open-source suites. *Science China Information Sciences*, 67(12):220101, 2024. doi: 10.1007/s11432-024-4231-5. URL https://doi.org/10.1007/s11432-024-4231-5.

Yunkai Dang, Meiyi Zhu, Donghao Wang, Yizhuo Zhang, Jiacheng Yang, Qi Fan, Yuekun Yang, Wenbin Li, Feng Miao, and Yang Gao. A Benchmark for Ultra-High-Resolution Remote Sensing MLLMs. *arXiv preprint arXiv:2512.17319v1*, 2025. URL https://arxiv.org/abs/2512.17319v1.

Yunkai Dang, Minxin Dai, Yuekun Yang, Zhangnan Li, Wenbin Li, Feng Miao, and Yang Gao. UHR-BAT: Budget-Aware Token Compression Vision-Language model for Ultra-High-Resolution Remote Sensing. In *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=AWPKhE1dbc.

Zonghao Guo, Ruyi Xu, Yuan Yao, Junbo Cui, Zanlin Ni, Chunjiang Ge, Tat-Seng Chua, Zhiyuan Liu, and Gao Huang. LLaVA-UHD: An LMM Perceiving Any Aspect Ratio and High-Resolution Images. In *Computer Vision – ECCV 2024*, volume 15141 of *Lecture Notes in Computer Science*, pp. 390–406. Springer, 2025. doi: 10.1007/978-3-031-73010-8_23. URL https://doi.org/10.1007/978-3-031-73010-8_23.

Yuan Hu, Jianlong Yuan, Congcong Wen, Xiaonan Lu, Yu Liu, and Xiang Li. RSGPT: A remote sensing vision language model and benchmark. *ISPRS Journal of Photogrammetry and Remote Sensing*, 224:272–286, 2025. doi: 10.1016/j.isprsjprs.2025.03.028. URL https://doi.org/10.1016/j.isprsjprs.2025.03.028.

Mingxin Huang, Yuliang Liu, Dingkang Liang, Lianwen Jin, and Xiang Bai. Mini-Monkey: Alleviating the Semantic Sawtooth Effect for Lightweight MLLMs via Complementary Image Pyramid. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=71XtUhazG0.

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. URL https://openaccess.thecvf.com/content/CVPR2024/html/Kuckreja_GeoChat_Grounded_Large_Vision-Language_Model_for_Remote_Sensing_CVPR_2024_paper.html.

Hugo Laurençon, Andrés Marafioti, Victor Sanh, and Léo Tronchon. Building and better understanding vision-language models: insights and future directions. *arXiv preprint arXiv:2408.12637v1*, 2024. URL https://arxiv.org/abs/2408.12637v1.

Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. LLaVA-OneVision: Easy Visual Task Transfer. *Transactions on Machine Learning Research*, 2025a. ISSN 2835-8856. URL https://openreview.net/forum?id=zKv8qULV6n.

Qingyun Li, Shuran Ma, Junwei Luo, Yi Yu, Yue Zhou, Fengxiang Wang, Xudong Lu, Xiaoxing Wang, Xin He, Yushi Chen, and Xue Yang. Co-Training Vision-Language Models for Remote Sensing Multi-Task Learning. *Remote Sensing*, 18(2):222, 2026. doi: 10.3390/rs18020222. URL https://doi.org/10.3390/rs18020222.

Wentong Li, Yuqian Yuan, Jian Liu, Dongqi Tang, Song Wang, Jie Qin, Jianke Zhu, and Lei Zhang. TokenPacker: Efficient Visual Projector for Multimodal LLM. *International Journal of Computer Vision*, 133(10):6794–6812, 2025b. doi: 10.1007/s11263-025-02491-7. URL https://doi.org/10.1007/s11263-025-02491-7.

Zhang Li, Biao Yang, Qiang Liu, Zhiyin Ma, Shuo Zhang, Jingxu Yang, Yabo Sun, Yuliang Liu, and Xiang Bai. Monkey: Image Resolution and Text Label Are Important Things for Large Multi-modal Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 26763–26773, 2024. URL https://openaccess.thecvf.com/content/CVPR2024/html/Li_Monkey_Image_Resolution_and_Text_Label_Are_Important_Things_for_CVPR_2024_paper.html.

Alexander H. Liu, Kartik Khandelwal, Sandeep Subramanian, Victor Jouault, Abhinav Rastogi, Adrien Sadé, Alan Jeffares, Albert Jiang, Alexandre Cahill, Alexandre Gavaudan, Alexandre Sablayrolles, Amélie Héliou, Amos You, Andy Ehrenberg, Andy Lo, Anton Eliseev, Antonia Calvi, Avinash Sooriyarachchi, Baptiste Bout, Baptiste Rozière, Baudouin De Monicault, Clémence Lanfranchi, Corentin Barreau, Cyprien Courtot, Daniele Grattarola, Darius Dabert, Diego de las Casas, Elliot Chane-Sane, Faruk Ahmed, Gabrielle Berrada, Gaëtan Ecrepont, Gauthier Guinet, Georgii Novikov, Guillaume Kunsch, Guillaume Lample, Guillaume Martin, Gunshi Gupta, Jan Ludziejewski, Jason Rute, Joachim Studnia, Jonas Amar, Joséphine Delas, Josselin Somerville Roberts, Karmesh Yadav, Khyathi Chandu, Kush Jain, Laurence Aitchison, Laurent Fainsin, Léonard Blier, Lingxiao Zhao, Louis Martin, Lucile Saulnier, Luyu Gao, Maarten Buyl, Margaret Jennings, Marie Pellat, Mark Prins, Mathieu Poirée, Mathilde Guillaumin, Matthieu Dinot, Matthieu Futeral, Maxime Darrin, Maximilian Augustin, Mia Chiquier, Michel Schimpf, Nathan Grinsztajn, Neha Gupta, Nikhil Raghuraman, Olivier Bousquet, Olivier Duchenne, Patricia Wang, Patrick von Platen, Paul Jacob, Paul Wambergue, Paula Kurylowicz, Pavankumar Reddy Muddireddy, Philomène Chagniot, Pierre Stock, Pravesh Agrawal, Quentin Torroba, Romain Sauvestre, Roman Soletskyi, Rupert Menneer, Sagar Vaze, Samuel Barry, Sanchit Gandhi, Siddhant Waghjale, Siddharth Gandhi, Soham Ghosh, Srijan Mishra, Sumukh Aithal, Szymon Antoniak, Teven Le Scao, Théo Cachet, Theo Simon Sorg, Thibaut Lavril, Thiziri Nait Saada, Thomas Chabal, Thomas Foubert, Thomas Robert, Thomas Wang, Tim Lawson, Tom Bewley, Tom Bewley, Tom Edwards, Umar Jamil, Umberto Tomasini, Valeriia Nemychnikova, Van Phung, Vincent Maladière, Virgile Richard, Wassim Bouaziz, Wen-Ding Li, William Marshall, Xinghui Li, Xinyu Yang, Yassine El Ouahidi, Yihan Wang, Yunhao Tang, and Zaccharie Ramzi. Ministral 3. *arXiv preprint arXiv:2601.08584v1*, 2026a. URL https://arxiv.org/abs/2601.08584v1.

Fan Liu, Delong Chen, Zhangqingyun Guan, Xiaocong Zhou, Jiale Zhu, Qiaolin Ye, Liyong Fu, and Jun Zhou. RemoteCLIP: A Vision Language Foundation Model for Remote Sensing. *IEEE Transactions on Geoscience and Remote Sensing*, 62:1–16, 2024. doi: 10.1109/tgrs.2024.3390838. URL https://doi.org/10.1109/tgrs.2024.3390838.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual Instruction Tuning. In *Advances in Neural Information Processing Systems*, volume 36, pp. 34892–34916, 2023. doi: 10.52202/075280-1516. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/6dcf277ea32ce3288914faf369fe6de0-Abstract-Conference.html.

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. *arXiv preprint arXiv:2509.22221v2*, 2026b. URL https://arxiv.org/abs/2509.22221v2.

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. *arXiv preprint arXiv:2511.12267v1*, 2025. URL https://arxiv.org/abs/2511.12267v1.

Gen Luo, Yiyi Zhou, Yuxin Zhang, Xiawu Zheng, Xiaoshuai Sun, and Rongrong Ji. Feast Your Eyes: Mixture-of-Resolution Adaptation for Multimodal Large Language Models. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=1EnpStvBU8.

Xianzhi Ma, Shujun Wang, Xiaohan Li, Hao Liu, Changhua Pei, and Jianhui Li. WeaveEarth: Structured Evidence Construction and Reasoning for Training-Free UHR Remote Sensing Understanding. *arXiv preprint arXiv:2607.10120v1*, 2026. URL https://arxiv.org/abs/2607.10120v1.

Zhongkuan Mao, Xianjie Liu, Tianyu Meng, Yidong Wang, Wenzhuo Zhao, Ronghao Xian, Yao Jiang, Fei Shen, Junfeng Fang, Yong Dai, Yi Zhang, and Keren Fu. Thinking Once Is Enough: Intermediate-Layer Evidence Routing for High-Resolution VQA. *arXiv preprint arXiv:2607.27830v1*, 2026. URL https://arxiv.org/abs/2607.27830v1.

Dilxat Muhtar, Zhenshi Li, Feng Gu, Xueliang Zhang, and Pengfeng Xiao. LHRS-Bot: Empowering Remote Sensing with VGI-Enhanced Large Multimodal Language Model. In *Computer Vision – ECCV 2024*, volume 15132 of *Lecture Notes in Computer Science*, pp. 440–457. Springer, 2025. doi: 10.1007/978-3-031-72904-1_26. URL https://doi.org/10.1007/978-3-031-72904-1_26.

Chao Pang, Xingxing Weng, Jiang Wu, Jiayu Li, Yi Liu, Jiaxing Sun, Weijia Li, Shuai Wang, Litong Feng, Gui-Song Xia, and Conghui He. 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. doi: 10.1609/aaai.v39i6.32683. URL https://doi.org/10.1609/aaai.v39i6.32683.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning Transferable Visual Models From Natural Language Supervision. In *International Conference on Machine Learning*, pp. 8748–8763, 2021. URL https://proceedings.mlr.press/v139/radford21a.html.

Hao Shao, Shengju Qian, Han Xiao, Guanglu Song, Zhuofan Zong, Letian Wang, Yu Liu, and Hongsheng Li. Visual CoT: Advancing Multi-Modal Language Models with a Comprehensive Dataset and Benchmark for Chain-of-Thought Reasoning. In *Advances in Neural Information Processing Systems*, volume 37, pp. 8612–8642, 2024. doi: 10.52202/079017-0275. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/0ff38d72a2e0aa6dbe42de83a17b2223-Abstract-Datasets_and_Benchmarks_Track.html.

Sagar Soni, Akshay Dudhane, Hiyam Debary, Mustansar Fiaz, Muhammad Akhtar Munir, Muhammad Sohail Danish, Paolo Fraccaro, Campbell D Watson, Levente J Klein, Fahad Shahbaz Khan, and Salman Khan. EarthDial: Turning Multi-sensory Earth Observations to Interactive Dialogues. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14303–14313, 2025. URL https://openaccess.thecvf.com/content/CVPR2025/html/Soni_EarthDial_Turning_Multi-sensory_Earth_Observations_to_Interactive_Dialogues_CVPR_2025_paper.html.

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. *arXiv preprint arXiv:2505.21375v2*, 2025a. URL https://arxiv.org/abs/2505.21375v2.

Fengxiang Wang, Hongzhen Wang, Zonghao Guo, Di Wang, Yulin Wang, Mingshuo Chen, Qiang Ma, Long Lan, Wenjing Yang, Jing Zhang, Zhiyuan Liu, and Maosong Sun. XLRS-Bench: Could Your Multimodal LLMs Understand Extremely Large Ultra-High-Resolution Remote Sensing Imagery? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14325–14336, 2025b. URL https://openaccess.thecvf.com/content/CVPR2025/html/Wang_XLRS-Bench_Could_Your_Multimodal_LLMs_Understand_Extremely_Large_Ultra-High-Resolution_Remote_CVPR_2025_paper.html.

Fengxiang Wang, Mingshuo Chen, Yueying Li, Yajie Yang, Yifan Zhang, Long Lan, Xue Yang, Hongda Sun, Yulin Wang, Di Wang, Jun Song, Jing Zhang, and Bo Du. GeoEyes: On-Demand Visual Focusing for Evidence-Grounded Understanding of Ultra-High-Resolution Remote Sensing Imagery. *arXiv preprint arXiv:2602.14201v2*, 2026a. URL https://arxiv.org/abs/2602.14201v2.

Fengxiang Wang, Jiangnan Huang, Mingshuo Chen, Yueying Li, Yang Shi, Junwei Luo, Haoyu Wang, Yansheng Li, Jing Zhang, Haiyan Zhao, and Wenjing Yang. Beyond Zooming: Learning Multi-Tool Visual Reasoning for Ultra-High-Resolution Remote Sensing. *arXiv preprint arXiv:2607.25993v1*, 2026b. URL https://arxiv.org/abs/2607.25993v1.

Penghao Wu and Saining Xie. V*: Guided Visual Search as a Core Mechanism in Multimodal LLMs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13084–13094, 2024. URL https://openaccess.thecvf.com/content/CVPR2024/html/Wu_V_Guided_Visual_Search_as_a_Core_Mechanism_in_Multimodal_CVPR_2024_paper.html.

Zhiyu Wu, Xiaokang Chen, Zizheng Pan, Xingchao Liu, Wen Liu, Damai Dai, Huazuo Gao, Yiyang Ma, Chengyue Wu, Bingxuan Wang, Zhenda Xie, Yu Wu, Kai Hu, Jiawei Wang, Yaofeng Sun, Yukun Li, Yishi Piao, Kang Guan, Aixin Liu, Xin Xie, Yuxiang You, Kai Dong, Xingkai Yu, Haowei Zhang, Liang Zhao, Yisong Wang, and Chong Ruan. DeepSeek-VL2: Mixture-of-Experts Vision-Language Models for Advanced Multimodal Understanding. *arXiv preprint arXiv:2412.10302v1*, 2024. URL https://arxiv.org/abs/2412.10302v1.

Jianwei Yang, Hao Zhang, Feng Li, Xueyan Zou, Chunyuan Li, and Jianfeng Gao. Set-of-Mark Prompting Unleashes Extraordinary Visual Grounding in GPT-4V. *arXiv preprint arXiv:2310.11441v2*, 2023. URL https://arxiv.org/abs/2310.11441v2.

Jiabo Ye, Anwen Hu, Haiyang Xu, Qinghao Ye, Ming Yan, Guohai Xu, Chenliang Li, Junfeng Tian, Qi Qian, Ji Zhang, Qin Jin, Liang He, Xin Lin, and Fei Huang. UReader: Universal OCR-free Visually-situated Language Understanding with Multimodal Large Language Model. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 2841–2858, 2023. doi: 10.18653/v1/2023.findings-emnlp.187. URL https://aclanthology.org/2023.findings-emnlp.187/.

Haoxuan You, Haotian Zhang, Zhe Gan, Xianzhi Du, Bowen Zhang, Zirui Wang, Liangliang Cao, Shih-Fu Chang, and Yinfei Yang. Ferret: Refer and Ground Anything Anywhere at Any Granularity. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=2msbbX3ydD.

Tianyu Yu, Zefan Wang, Chongyi Wang, Fuwei Huang, Wenshuo Ma, Zhihui He, Tianchi Cai, Weize Chen, Yuxiang Huang, Ranchi Zhao, Bokai Xu, Junbo Cui, Yingjing Xu, Liqing Ruan, Luoyuan Zhang, Hanyu Liu, Jingkun Tang, Hongyuan Liu, Qining Guo, Wenhao Hu, Bingxiang He, Jie Zhou, Jie Cai, Ji Qi, Zonghao Guo, Chi Chen, Guoyang Zeng, Yuxuan Li, Ganqu Cui, Ning Ding, Xu Han, Yuan Yao, Zhiyuan Liu, and Maosong Sun. MiniCPM-V 4.5: Cooking Efficient MLLMs via Architecture, Data, and Training Recipe. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11704–11715, 2026. URL https://openaccess.thecvf.com/content/CVPR2026/html/Yu_MiniCPM-V_4.5_Cooking_Efficient_MLLMs_via_Architecture_Data_and_Training_CVPR_2026_paper.html.

Wei Zhang, Miaoxin Cai, Tong Zhang, Yin Zhuang, and Xuerui Mao. EarthGPT: A Universal Multimodal Large Language Model for Multisensor Image Comprehension in Remote Sensing Domain. *IEEE Transactions on Geoscience and Remote Sensing*, 62:1–20, 2024. doi: 10.1109/tgrs.2024.3409624. URL https://doi.org/10.1109/tgrs.2024.3409624.

Zilun Zhang, Haozhan Shen, Tiancheng Zhao, Zian Guan, Bin Chen, Yuhao Wang, Xu Jia, Yuxiang Cai, Yongheng Shang, and Jianwei Yin. Enhancing Ultrahigh Resolution Remote Sensing Imagery Analysis With ImageRAG: A new framework. *IEEE Geoscience and Remote Sensing Magazine*, 13(3):369–394, 2025. doi: 10.1109/mgrs.2025.3574742. URL https://doi.org/10.1109/mgrs.2025.3574742.

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:2511.20460v2*, 2025. URL https://arxiv.org/abs/2511.20460v2.

Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, Zhangwei Gao, Erfei Cui, Xuehui Wang, Yue Cao, Yangzhou Liu, Xingguang Wei, Hongjie Zhang, Haomin Wang, Weiye Xu, Hao Li, Jiahao Wang, Nianchen Deng, Songze Li, Yinan He, Tan Jiang, Jiapeng Luo, Yi Wang, Conghui He, Botian Shi, Xingcheng Zhang, Wenqi Shao, Junjun He, Yingtong Xiong, Wenwen Qu, Peng Sun, Penglong Jiao, Han Lv, Lijun Wu, Kaipeng Zhang, Huipeng Deng, Jiaye Ge, Kai Chen, Limin Wang, Min Dou, Lewei Lu, Xizhou Zhu, Tong Lu, Dahua Lin, Yu Qiao, Jifeng Dai, and Wenhai Wang. InternVL3: Exploring Advanced Training and Test-Time Recipes for Open-Source Multimodal Models. *arXiv preprint arXiv:2504.10479v3*, 2025. URL https://arxiv.org/abs/2504.10479v3.

A Benchmark Characteristics and Evaluation Protocol

A.1 Task taxonomy and evaluation support

Tables S1 and S2 summarize the evaluated support: XLRS-Bench contains 3,080 questions across 13 tasks, while our fixed RSHR single-turn subset contains 1,590 questions across 14 tasks. The RSHR counts refer only to this subset; its additional multi-turn tasks are not evaluated (Dang et al., 2025). All reported tasks, including those named grounding or detection, use the benchmark VQA answer interface. Native and GeoEPS receive the same original images, questions, answer choices, and spatial cues.

Table S1: **Task characteristics of XLRS-Bench.** The two task groups cover perception and reasoning in UHR-RS VQA (Wang et al., 2025b). Counts are questions, not unique images. OLUC uses permitted option sets; OMS uses A/B for yes/no.

L1 task	L2 task	L3 task	Code	N	Answer format
Perception	Counting	Overall counting	OC	60	A/B/C/D
		Regional counting	RC	100	A/B/C/D
	Scene classification	Overall land-use classification	OLUC	100	Option set
		Regional land-use classification	RLUC	200	A/B/C/D
	Object spatial relationship	Object spatial relationships	OSR	500	A/B/C/D
	Object properties	Object classification	OCC	800	A/B/C/D
		Object color	OCL	800	A/B/C/D
		Object motion state	OMS	60	A/B
Reasoning	Route planning	Route planning	RP	100	A/B/C/D
	Anomaly reasoning	Anomaly detection and interpretation	AD	100	A/B/C/D
	Complex reasoning	Environmental condition reasoning	ECR	100	A/B/C/D
		Counting with complex reasoning	CCR	100	A/B/C/D
	Spatiotemporal reasoning	Regional counting with change detection	RCCD	60	A/B/C/D
Total: 13 tasks				**3,080**	

Matched evaluation support. GeoChat excludes RCCD on XLRS and MRJC on RSHR, yielding 3,020 questions across 12 tasks and 1,570 questions across 13 tasks, respectively. Each Native/GeoEPS pair uses the same evaluated questions and tasks. XLRS OC denotes overall counting, whereas RSHR OC denotes object classification.

Cue routing. GeoEPS uses EAP for accepted numeric pointers or visual markers and CES for no-cue inputs, subject to the input-support exceptions below. This choice is determined by the supplied cues, not task labels.

XLRS has 1,660 numeric-pointer questions (OCC/OCL/OMS: 800/800/60) and 346 visual-marker questions (RC/RLUC/RCCD: 94/194/58). These 2,006 questions use EAP. Of the remaining 1,074 no-cue questions, 1,072 use CES, including six RC and six RLUC questions. The other CES tasks are OC, OLUC, OSR, AD, ECR, RP, and CCR. The two no-cue multi-image RCCD questions use Native predictions.

On RSHR, 700 numeric-pointer questions use EAP (COL/SHP/DET/OC/RCN: 200/100/100/200/100). For Qwen3, CES processes 820 no-cue questions with one source image. The remaining 50 two-image and 20 three-image questions use predictions from their multi-image evaluation configurations and are outside the single-source CES procedure. ZoomEarth, GeoLens, and GeoEyes use their native evidence acquisition outside EAP, as denoted by * in the main tables.

Table S2: **Task characteristics of the fixed RSHR-Bench subset.** Names follow the benchmark taxonomy (Dang et al., 2025); counts refer only to the 1,590 questions used in Table 2. Each response selects one of the listed option labels.

L1 task	Task family	Evaluated task	Code	N	Answer format
Perception	Local attributes	Color	COL	200	A/B/C/D
		Shape	SHP	100	A/B/C/D
		Orientation detection	DET	100	A/B/C/D
	Overall attributes	Object classification	OC	200	A/B/C/D
		Object spatial relationships	REL	200	A/B/C/D
	Visual grounding	Object grounding	OGD	200	A/B/C/D
		Regional grounding	RG	70	A/B/C/D
	Counting	Object counting	OCN	200	A/B/C/D
		Regional counting	RCN	100	A/B/C/D
Reasoning	Knowledge-grounded	Anomaly detection	AD	50	A/B/C/D
		Future prediction	FP	50	A/B/C/D
	Logic-grounded	Multi-region joint contrast (multiple images)	MRJC	20	A/B/C
		Multi-region joint contrast (single image)	MRJCS	50	A/B/C
		Object-state judgment	OSJ	50	A/B/C/D
Total: 14 tasks				**1,590**	

A.2 Macro accuracy and answer parsing

For an evaluated task set $\mathcal{T}$, let n_t be the number of questions in task t and k_t the number whose parsed responses match the benchmark answers. The aggregate accuracy reported in this appendix is

$$\text{Macro} = \frac{100}{|\mathcal{T}|} \sum_{t \in \mathcal{T}} \frac{k_t}{n_t}. \tag{11}$$

Macro gives each evaluated task equal weight. On XLRS, Macro_3 averages OCC, OCL, and OMS; Macro_9 averages AD, CCR, ECR, OC, OLUC, OSR, RC, RLUC, and RP. Macro_9 uses the full question support of these nine tasks, including the cue-bearing RC and RLUC questions, as detailed in Appendix D.2. RSHR Macro averages the evaluated individual tasks with equal weight. Changes in Macro are reported in percentage points (pp). Accuracy tables use one decimal place, and displayed changes are computed from the displayed accuracies.

Single-choice responses are parsed as option letters. Permitted option sets, including XLRS OLUC answers, are normalized to uppercase, distinct, alphabetically ordered letters and scored by exact set equality. Unresolved or disallowed responses receive zero credit and remain in the denominator. For CES, a parsed hypothesis in $\mathcal{Y}(q)$ is eligible based on its permitted answer format, independently of its agreement with the benchmark answer.

A.3 Inference settings and answer-interface scope

Table S3 summarizes the Qwen3 inference settings. The detailed input controls, stage-wise comparisons, and cue-perturbation experiments in this appendix are conducted on XLRS. The same frozen backbone answers and verifies; no benchmark answer label enters an inference call.

Candidate answering uses the benchmark question-and-choice interface and parser, with the query alignment specified for EAP. CES returns the answer associated with the selected evidence candidate without an additional answer-generation call, as specified in Section 3.3 and Appendix C. Other backbones use their model-specific input processors and decoding configurations.

Table S3: **Qwen3 inference settings.** Image-size limits specify the minimum and maximum pixel counts ($H \times W$) used by the image processor for each image.

Setting	Configuration
Backbone	Qwen3-VL-8B-Instruct.
EAP image-size limits	200,704–1,605,632 pixels per image.
CES overview limits	65,536–1,048,576 pixels per overview image.
CES local view	Source-image crop, passed to the image processor without a separate resize.
XLRS candidate answer	Greedy decoding; at most 64 new tokens.
RSHR candidate answer	Greedy decoding; at most 128 new tokens.
Verification	YES/NO token scores; prompts and scoring are specified in Appendix C.1.

B Evidence-Aware Presentation: Construction and Controls

EAP combines contextual cropping, textual reference alignment, visual marking, and query-conditioned presentation. This section gives the construction rules and compares their effects across query families. All controls below use the same 1,660 XLRS numeric-pointer questions with frozen Qwen3-VL-8B. OCC, OCL, and OMS denote object classification, object color, and object motion state, respectively. Macro_3 is their unweighted mean accuracy.

B.1 Cue parsing and contextual-view construction

The router checks for a numeric pointer first, an existing visual marker second, and otherwise follows the no-cue protocol in Appendix A: CES is used for supported single-source inputs, while the two no-cue multi-image XLRS questions use Native predictions. Table S4 records the fixed geometric and rendering choices; constructing an input does not consult its ground-truth answer. The query-conditioned selector applies only to numeric pointers.

Table S4: **EAP construction settings.** Crops and marks are constructed before the VLM's image processor. κ scales the mapped reference box.

Component	Fixed rule
Numeric reference	Adjacent image-resolution and bounding-box fields in the public query; integer pixel corners in the stated frame.
Context expansion	$\kappa = 4$ for numeric pointers; $\kappa = 2$ for existing markers; minimum side length 64 pixels.
Context boundaries	Clip each endpoint and round half to even; no padding or inward translation.
Outside outline	RGB $(255, 0, 0)$; width 4 native-crop pixels; no fill or antialiasing.
Existing red marker	$R \geq 150$ and $R - \max(G, B) \geq 80$; eight-connected components.
Marker acceptance	At most six components in decreasing area; area ≥ 500 pixels and added-hole-fill area / component area ≥ 0.5.
Marker view	Enclosing rectangle of accepted components plus context; preserve marker and query.
Multiple source images	Preserve order; each must contain an accepted marker. V_{prov} then denotes the ordered contextual views.

A nonpositive image size, a box degenerate after mapping/clipping, or multiple complete numeric specifications does not activate the pointer route. Crop rectangles are half-open: left/top endpoints are included and right/bottom endpoints are excluded.

B.2 Contextual crop and outside outline

Let (W_s, H_s) be the image dimensions stated in the query and (W_a, H_a) those of the corresponding input image. Map the supplied box componentwise using W_a/W_s and H_a/H_s, then clip to the input-image boundaries. For the mapped box, with center (c_x, c_y) and width and height (w, h), the

contextual width w_c, height h_c, and requested box k are

$$w_c = \max(\kappa w, 64), \qquad h_c = \max(\kappa h, 64),$$
$$k = (c_x - w_c/2, c_y - h_c/2, c_x + w_c/2, c_y + h_c/2), \tag{12}$$

where $\kappa = 4.0$ for a numeric pointer and 2.0 for an existing marker. The minimum side length is 64 pixels before clipping. The requested window is clipped componentwise and then rounded with Python `int(round(v))`, which uses half-to-even rounding. It is not shifted back inside the image or padded; context can therefore be asymmetric at edges.

For the numeric reference within the crop, subtract the clipped but *unrounded* crop origin from the mapped reference, then round and clamp the local box to a nondegenerate rectangle. The resulting local box is $\widetilde{g}$ in Eq. 2. The outside band is painted in native contextual-crop pixels before the VLM's image processor: RGB $(255, 0, 0)$, width four, no fill and no antialiasing. Its mask is the four-pixel-expanded box minus the reference-box interior, including the outer corner pixels. A band crossing the crop boundary is truncated, not moved inward.

B.3 Query templates and presentation selection

The numeric reference specification is replaced once with one of these literal phrases; answer choices remain in their original order.

> **Marked reference:** The referenced object is the object outlined by the red rectangle.
> **Clean reference:** The referenced object is the object indicated by the local crop.

JOINT uses the marked-reference wording above and supplies the marked crop first and its clean counterpart second. Its role prefix is

> Two views of the same referenced region are provided.
> The first view marks the referenced object or region with a red rectangle.
> The second view is the same local crop without the annotation.
> Use the marked view to identify the referent and the clean view to inspect its visual evidence.

The prefix is followed by a newline and the aligned question body, followed by the original answer choices. Their order and the benchmark answer-format instruction are preserved.

This paired presentation supplies an explicit referent cue together with an annotation-free view of the same region. For color queries, the intended division of roles is to locate the referenced object in V_{mark} and inspect its color in V_{clean}. Both views have the same crop geometry and surrounding context; the clean counterpart contains no added outline.

Query-conditioned selection. At inference, the fixed selector $c(\cdot)$ evaluates textual patterns on the original query, not answer choices, task IDs, pixels, or correctness labels. For XLRS, the selector uses the following three fixed query-matching patterns:

> Color → JOINT (case-insensitive):
> `determine\s+the\s+colou?r\s+of\s+the\s+object`
>
> Category → REFERENCE (case-insensitive):
> `identify\s+the\s+object\s+category`
>
> Motion → CLEAN (case-insensitive, dot matches newlines):
> `determine\s+whether\s+the\s+object\b.*?\bis\s+in\s+motion`

All patterns are evaluated; a single match determines the mode, while zero or multiple matches select JOINT. Every evaluated XLRS pointer query matches exactly one family. Table S5 summarizes the XLRS presentations. Existing-marker inputs use the original query unchanged and bypass the pointer-specific selector.

Table S5: **Query-conditioned presentations on XLRS.** Counts cover all 1,660 pointer questions; the mode names denote input presentations, not separate models.

Query family	Mode	Images	Aligned query	N
Object category	REFERENCE	V_{mark}	q_{mark}	800
Motion state	CLEAN	V_{clean}	q_{clean}	60
Color	JOINT	$[V_{\text{mark}}, V_{\text{clean}}]$	q_{joint}	800

B.4 Fixed versus query-conditioned presentations

Table S6: **Fixed and query-conditioned EAP presentations** on 1,660 XLRS pointer questions with Qwen3-VL-8B. REFERENCE/CLEAN/JOINT use their associated aligned queries. All accuracy columns are percentages; Images/q is the mean number of supplied images, not a measure of runtime. Bold marks the highest accuracy in each column.

Presentation	OCC	OCL	OMS	Macro_3	Images/q
Fixed REFERENCE	92.5	86.5	78.3	85.8	1
Fixed CLEAN	79.8	60.8	**90.0**	76.8	1
Fixed JOINT	**93.1**	**88.0**	85.0	88.7	2
EAP	92.5	**88.0**	**90.0**	**90.2**	1.48

Table S6 shows complementary strengths: CLEAN has the highest OMS accuracy, while JOINT improves OCL over either single view. The OCL result is consistent with the intended combination of explicit referent identification and annotation-free appearance inspection in JOINT. These comparisons evaluate the complete presentations, including their aligned queries and, for JOINT, the role prefix. EAP achieves the highest Macro_3 with $1 + 800/1660 \approx 1.48$ images per question, compared with two for fixed JOINT. Fixed JOINT has slightly higher OCC accuracy, while EAP has higher task-balanced accuracy.

B.5 Progressive construction and coordinate-correct cropping

Table S7 complements the main construction table with taskwise results and Macro_3 gains from E0. E0 uses the original image and query. E1 uses a tight crop with the original query. The E2 variants share that tight crop: E2-R remaps the numeric reference into crop coordinates, whereas E2 replaces it with an aligned local-crop reference. E3 adds context. The E4 variants use a marked reference: E4-C centers a four-pixel outline on the box boundary (two pixels inside and two outside), while E4 places it entirely outside. E5 selects REFERENCE, CLEAN, or JOINT by query; its outline is absent for CLEAN and present in the marked view of the other modes. E2-R and E2 are alternative reference controls relative to E1; E4-C and E4 are alternative marking controls relative to E3. The stepwise gains in Table 4 use these shared bases, rather than the immediately preceding row for each alternative. Both E3-to-E4-C and E3-to-E4 change reference text as well as rendering; they do not isolate a pixel-only marking effect. The matched rendering control below holds the reference text fixed to isolate that effect.

EAP exceeds E2-R by 16.1 Macro_3 points. Adding context at E3 improves OMS but reduces OCC and OCL relative to E2, for a net gain of 1.0 Macro_3 points. With the same contextual crop, the marked-reference presentation at E4 raises OCL from 60.8% to 86.5%. These results support combining context with explicit reference presentation.

B.6 Matched rendering and view-composition controls

Let q_N denote a marker-neutral query using the numeric reference. The R0/R1/R2 controls use identical contextual crops, answer choices, processing, and the same q_N. This text is generated per question by updating `Image resolution:` to crop dimensions and `Bounding box:` to rounded local coordinates; each question therefore has its own coordinate-correct reference. R0 is clean, R1 adds the outside outline and a 20%-opacity interior fill, and R2 uses only the outline.

Table S7: **EAP construction controls with Qwen3-VL-8B.** OCC/OCL/OMS report accuracies on 800/800/60 questions, respectively. All accuracy columns are percentages; Δ_0 is the Macro_3 gain over E0 (pp), computed from the displayed Macro_3 values. The IDs match Table 4.

ID	OCC	OCL	OMS	Macro_3	Δ_0
E0	46.6	41.1	66.7	51.5	–
E1	77.9	65.3	65.0	69.4	+17.9
E2-R	80.3	78.8	63.3	74.1	+22.6
E2	85.6	78.4	63.3	75.8	+24.3
E3	79.8	60.8	90.0	76.8	+25.3
E4-C	91.9	86.6	71.7	83.4	+31.9
E4	92.5	86.5	78.3	85.8	+34.3
E5	92.5	88.0	90.0	90.2	+38.7

Table S8: **Matched rendering controls** with Qwen3-VL-8B. Only rendering changes; all three conditions use the same per-question numeric reference text q_N.

Rendering	OCC	OCL	OMS	Macro_3
R0: clean	59.0	39.3	73.3	57.2
R1: fill + outline	57.9	40.5	66.7	55.0
R2: outline only	87.1	79.1	71.7	79.3

The outline improves Macro_3 by 22.1 points over clean input and 24.3 points over the filled rendering. This supports the unfilled design under the tested crop and text settings, without attributing the gain to query rewriting. The fill is a diagnostic intervention, not a deployed EAP mode.

A separate control holds the two-image prefix, query, image count, crop dimensions, and processor settings fixed while changing which crop versions occupy the slots.

Table S9: **Two-image composition controls** with Qwen3-VL-8B. Both slots remain present and contain versions of the same contextual crop. The common neutral prefix differs from the deployed JOINT prompt.

Composition	OCC	OCL	OMS	Macro_3
marked + clean	89.4	81.9	80.0	83.8
marked + marked	90.4	83.8	66.7	80.3
clean + clean	68.1	50.4	65.0	61.2

Marked+clean has the highest Macro_3. Compared with marked+marked, it improves OMS but slightly reduces OCC and OCL. Thus, view composition matters even when image count, crop dimensions, and processor settings are fixed. This control compares two-image compositions under a neutral prefix; Table S6 compares the deployed presentations with their aligned queries and the JOINT role prefix. Because the image count is fixed here, this control does not isolate the effect of adding a second image.

C Competitive Evidence Selection: Search and Evidence Comparison

CES searches for evidence when the input provides no explicit spatial cue. It ranks local views by self-verification, then compares the top-ranked view (the incumbent) with a different view that produces a different answer hypothesis (the challenger). This section specifies the global-local inputs, crop geometry, verifier templates, and deterministic selection rules for the Qwen3 implementation on single-source XLRS inputs. The notation follows Section 3.3; input-support exceptions are given in Appendix A.

C.1 GLOBAL-LOCAL INPUTS AND VERIFIER INSTRUCTIONS

The Qwen3 implementation of the CES module searches one source image. For an evidence candidate e, $\Pi(e)$ is assembled as four ordered content blocks: the source image for the overview, footprint text, local crop, and a final text block for answering or verification. The local crop is extracted from the same source image as the overview, without a separate resize before the VLM's image processor. The footprint text is

```
Local evidence 1.  Source-scene footprint
<footprint> in normalized 0-1000 coordinates:
```

Here `<footprint>` is replaced by the candidate's coordinate string; it is not literal prompt text. The coordinates locate the local crop within the source image.

Candidate answering. Candidate answers use the benchmark answering interface and parser, with greedy decoding and at most 64 new tokens on XLRS. The question is followed by the released choice text, preserving option order and the benchmark answer-format instruction. Parsed answers are normalized to allowed letters or option sets and are returned directly if their candidate is selected. Verifier calls separate the query, choices, and SV or DV instruction with newlines.

Self-verification. The SV template $\mathcal{T}_{\mathrm{SV}}$ appends the literal instruction

> Based only on the visual evidence visible in this image, is there enough relevant visual information to determine the correct answer to the question? Answer YES or NO.

The line wrapping above is typographic. The response words `YES` and `NO` each correspond to one token. The YES next-token log-probability minus the NO next-token log-probability gives $s(e)$ in Eq. 6. This score measures model-assessed evidence sufficiency; it is not a calibrated probability of answer correctness.

Discriminative verification. The DV template $\mathcal{T}_{\mathrm{DV}}$ substitutes the ordered hypotheses into the literal instruction

> Does this image contain visual evidence that distinguishes whether the correct answer is `%s` or `%s` for the given question?
>
> Answer YES or NO.

The blank line is part of this instruction. The substitutions are the incumbent and challenger hypotheses, h_1 and h_2: normalized option letters or permitted option sets, not option descriptions. Both candidates receive the same ordered pair and final text block; their local views and footprint text differ. DV uses the same YES/NO scoring rule as SV, giving $D_F(\Pi(e); h_1, h_2)$ in Eq. 8.

C.2 SEARCH SCHEDULE AND CANDIDATE CONSTRUCTION

Table S10 summarizes the fixed search schedule: five canonical views cover different parts of the image, the two highest-scoring views undergo scale refinement, and CROSS5 shifts the best refined view in four directions. The listed SV calls exclude candidate-answer generation and the two DV evaluations performed when a challenger exists.

Table S10: **Sparse CES search schedule.** Reused scores do not require a new SV forward. Candidate geometry and enumeration are fixed during inference.

Stage	Evidence candidates	Ranked	SV calls
Canonical	Four corners and center; 1.5× grid-cell extent	5	5
Scale refinement	Top two at 1.00× and 0.75×	4	2
CROSS5	Best refined window and four 20% shifts	5	4
Total search-stage SV calls			**11**

Canonical windows. The four corners and center of a 3×3 grid are enumerated as T11, T13, T22, T31, and T33, where the digits denote one-based row and column. For source width W and zero-based column j, define $a_0 = \lfloor jW/3 \rfloor$ and $a_1 = \lfloor (j+1)W/3 \rfloor$. The expanded window has

width u and horizontal endpoints

$$u = \min\big(W, \lceil 3(a_1 - a_0)/2 \rceil\big),$$
$$x_0 = \mathrm{clamp}\big(\lfloor (a_0 + a_1 - u)/2 \rfloor, 0, W - u\big), \qquad x_1 = x_0 + u, \tag{13}$$

where $\mathrm{clamp}(z, L, U) = \min(\max(z, L), U)$. The vertical construction uses the image height and grid row analogously. Clamping shifts the window into the image while preserving its size. Exact SV ties follow the enumeration above.

Scale refinement. The top two canonical candidates are refined at scales 1.00 and 0.75, in the order $e_1^{\mathrm{r}}, \ldots, e_4^{\mathrm{r}}$ defined in Section 3.3. Unit-scale views reuse their SV scores. For a parent window of width w, the smaller width is $\max(1, \lceil 3w/4 \rceil)$. Its origin is computed by floor division about the same doubled center, as in the canonical construction; height is treated analogously. The resulting integer rectangle is intersected with the source image. Exact ties keep the earliest candidate in the refinement order.

CROSS5 translation and candidate identity. The best refined window anchors CROSS5. Its execution order is

$$[e_{\mathrm{t}}^{\mathrm{r}}, e_{\mathrm{l}}^{\mathrm{r}}, e_{\mathrm{o}}^{\mathrm{r}}, e_{\mathrm{r}}^{\mathrm{r}}, e_{\mathrm{b}}^{\mathrm{r}}], \tag{14}$$

corresponding to top, left, center, right, and bottom. Displacements are 20% of the crop's width or height, converted to integers using Python `round`. The translated origin is clamped to $[0, W - w] \times [0, H - h]$, preserving the crop size (w, h). Exact SV ties follow the execution order above. Reusing the center score gives $5 + 2 + 4 = 11$ SV calls across the three search stages.

Boundary clamping can produce duplicate realized crops, which remain in the ranked pool. In this single-source search, evidence identity is

$$\mathrm{iid}(e) = (\text{integer crop rectangle}, \text{image-processing configuration}).$$

The image-processing configuration is fixed across these candidates; thus, identical crop rectangles from the same source have the same identity. A candidate sharing the incumbent's identity is skipped without generating another hypothesis. Among the remaining candidates, CES accepts the first in descending SV order whose parsed answer is valid and differs from the incumbent's.

C.3 Evidence selection and answer reuse

Table S11 summarizes the selection procedure after CROSS5 ranking. The same official answering interface produces candidate hypotheses and the returned prediction. The ordered pair (h_1, h_2) is held fixed across the two DV evaluations.

Table S11: **CES decision procedure.** The scan stops at the first candidate with a different evidence input and a different allowed answer. No answer is regenerated after selection.

Step	Operation
1	Obtain $h_1 = \mathrm{Ans}_F(\Pi(e_1))$.
2	Scan $e_2, \ldots, e_5$ in descending $s(e)$ order; skip $\mathrm{iid}(e) = \mathrm{iid}(e_1)$.
3	Obtain $h_2 = \mathrm{Ans}_F(\Pi(e))$ for each remaining candidate in turn. Accept the first allowed $h_2 \neq h_1$ as the challenger.
4	With no challenger, set $e_{\mathrm{sel}} = e_1$ and return $\hat{y} = h_1$.
5	Otherwise compute D_{inc} and D_{ch} under the same (h_1, h_2).
6	If $D_{\mathrm{ch}} > D_{\mathrm{inc}}$, select e_{ch} and return h_2; otherwise select e_1 and return h_1.

Unparseable alternatives are skipped. Any unresolved returned prediction receives zero credit under the benchmark parser. These parsing safeguards do not consult benchmark answer labels during inference. CES returns the hypothesis already associated with the selected candidate, without regenerating an answer.

D ADDITIONAL ABLATION DETAILS

This section specifies the evaluation support and controlled comparisons underlying the XLRS ablations in the main paper. The detailed EAP input controls and CES implementation are given in Appendix B and Appendix C, respectively.

D.1 SCOPE OF THE MODULE COMPARISON

In the XLRS module comparison in Table 3, EAP-only applies evidence-aware presentation to cue-bearing inputs and Native inference to no-cue inputs. Full GeoEPS additionally applies CES to the supported no-cue questions, while the EAP predictions remain fixed. The two no-cue multi-image RCCD questions use Native predictions, as specified in Appendix A. GeoEyes* uses its own evidence-acquisition procedure, and Table 3 does not report a separate CES-added configuration.

D.2 STAGE-WISE CES EVALUATION SUPPORT

Table 5 averages nine complete XLRS tasks containing 1,360 questions: 1,072 are processed by CES and 288 cue-bearing RC/RLUC questions use the same fixed EAP predictions. For each stage, each task's accuracy is computed after combining its CES-supported and fixed EAP predictions. Macro_9 is the unweighted mean of the nine resulting task accuracies. Thus, it measures performance over the full task support, including the fixed cue-bearing questions.

C0 uses Native inference on the 1,072 no-cue questions and uses the same fixed EAP outputs on the 288 cue-bearing questions. Its Native label therefore refers to the CES-supported portion, not to replacing every prediction in these nine tasks. Table S12 gives the evaluation support. The two multi-image RCCD exceptions are outside these nine tasks.

Table S12: **Evaluation support for the CES stage-wise ablation.** Entries count evaluated questions, not successful predictions. The cue-bearing EAP predictions are fixed across C0–C4.

Task	Full support	CES-supported	Fixed EAP
AD	100	100	0
CCR	100	100	0
ECR	100	100	0
OC	60	60	0
OLUC	100	100	0
OSR	500	500	0
RC	100	6	94
RLUC	200	6	194
RP	100	100	0
Total	1,360	1,072	288

D.3 SEARCH STAGES AND COMPETITIVE SELECTION

C1, C2, C3, and C4 answer from e_1^{s}, $e_{\mathrm{o}}^{\mathrm{r}}$, e_1, and e_{sel}, respectively, using the same benchmark answering interface. These correspond to canonical search, scale refinement, CROSS5 translation, and competitive selection in Section 3.3.

C3 and C4 share the CROSS5 candidate pool and SV ranking. C3 returns the top-SV incumbent; C4 additionally searches the ranked alternatives for the first identity-distinct candidate with a different allowed answer. When such a challenger exists, DV evaluates both candidates under the same ordered hypothesis pair. The challenger is selected only if its DV score is strictly higher; otherwise C4 selects the incumbent. C4 also selects the incumbent when no eligible challenger exists. The answer associated with the selected candidate is returned without regeneration.

With the candidate pool and SV ranking fixed, C4 improves Macro_9 from 52.9% to 53.4%, a gain of 0.5 percentage points over C3 (Table 5). This comparison measures the effect of competitive selection on the evaluated tasks.

E SENSITIVITY TO NUMERIC-POINTER PERTURBATIONS

E.1 PERTURBATION PROTOCOL

We test numeric-pointer imprecision on 1,660 XLRS questions (OCC/OCL/OMS: 800/800/60) with Qwen3-VL-8B-Instruct. Only the supplied box changes: the source image, answer choices, and evaluation answer for the originally intended target stay fixed. Translation shifts the box by 10%, 20%, 30%, or 50% of its own width or height. Scaling changes both box dimensions by $1.1\times$, $1.2\times$, $1.5\times$, or $2.0\times$ about the original center. Translation and scaling are evaluated separately, not cumulatively.

Let $b = (x_1, y_1, x_2, y_2)$ have width $w = x_2 - x_1$, height $h = y_2 - y_1$, and center (c_x, c_y). For translation fraction τ and unit direction $d = (d_x, d_y) \in \{(-1, 0), (1, 0), (0, -1), (0, 1)\}$, the perturbed box is

$$\delta_x = \mathrm{round}(\tau d_x w), \quad \delta_y = \mathrm{round}(\tau d_y h), \quad b' = b + (\delta_x, \delta_y, \delta_x, \delta_y). \tag{15}$$

For scale factor λ, it is

$$b' = \mathrm{round}(c_x - \lambda w/2, c_y - \lambda h/2, c_x + \lambda w/2, c_y + \lambda h/2). \tag{16}$$

Rounding is componentwise to integers using Python's half-to-even rule. Perturbations occur in the query's stated coordinate frame, before mapping to the input image. Endpoints are clipped to the stated image boundaries without translating the box inward or restoring target coverage. Clipping can therefore change the realized center of a nominally center-fixed scale perturbation.

Matched directions. Each question uses one fixed translation direction, shared by E2-R and GeoEPS at every magnitude. Within each task, questions are sorted by `SHA256(seed + "|direction|" + qid)`, with ties broken by qid, then assigned cyclically to LEFT, RIGHT, TOP, and BOTTOM. The seed is `GEOEPS_FINAL_ROBUSTNESS_QWEN3_XLRS_V1`. Each direction contains 200 OCC, 200 OCL, and 15 OMS questions, totaling 415. The score thus uses one matched direction per question, rather than a four-direction average or a worst-case perturbation.

E.2 COORDINATE-CORRECT CROPPING AND PRESENTATION

E2-R tightly crops the perturbed box and remaps its numeric reference to crop coordinates, preserving the rest of the query. The EAP module derives its contextual crop, outline, and aligned query from the same perturbed box using its unchanged presentation rule. The original clean box is not provided as an additional hint. On these numeric-pointer questions, GeoEPS uses its EAP branch; CES is not activated by this experiment.

E.3 RESULTS UNDER CUE IMPRECISION

Table 6 in the main paper reports GeoEPS and its change from the clean Macro_3 score. Table S13 adds the E2-R baseline and each method's change from its own clean score. Accuracies use one decimal place, matching the main paper; all changes are computed from these displayed values.

Table S13: **Robustness to numeric-pointer imprecision.** All scores are Macro_3 (%) on the same 1,660 XLRS questions. Each translation level uses one matched direction per question. Clean denotes the unperturbed reference. Δ_{clean} is the change in percentage points from each method's displayed clean score (74.1 for E2-R and 90.2 for GeoEPS). Bold marks the higher accuracy in each row.

		E2-R		GeoEPS	
Perturbation	Level	Macro_3	Δ_{clean}	Macro_3	Δ_{clean}
Clean	–	74.1	–	**90.2**	–
Translation	10%	72.8	−1.3	**88.8**	−1.4
	20%	71.0	−3.1	**88.6**	−1.6
	30%	69.2	−4.9	**87.5**	−2.7
	50%	64.2	−9.9	**84.4**	−5.8
Scale	1.1×	73.5	−0.6	**88.1**	−2.1
	1.2×	71.0	−3.1	**88.8**	−1.4
	1.5×	67.4	−6.7	**88.1**	−2.1
	2.0×	63.5	−10.6	**85.3**	−4.9

GeoEPS has higher Macro_3 than E2-R throughout the tested range, although its decrease from clean is not always smaller. At 50% translation, GeoEPS decreases by 5.8 points to 84.4%, compared with a 9.9-point decrease to 64.2% for E2-R. At 2.0× scale, the decreases are 4.9 and 10.6 points, respectively.

These results show tolerance to the tested translations and scaling when answers are scored against the originally intended target. A shifted reference can still omit that target or point to another object, so these results do not imply recovery from arbitrary reference errors.

F Qualitative Analysis

F.1 Complementary marked and clean views

Figures S1 and S2 show four recorded Qwen3-VL-8B numeric-pointer examples from XLRS, comparing Native, REFERENCE, CLEAN, JOINT, and query-conditioned EAP. Each presentation mode uses its aligned query (Appendix B.3). EAP selects the mode from the original query text. These cases illustrate how predictions change with presentation; they do not measure how often each outcome occurs.

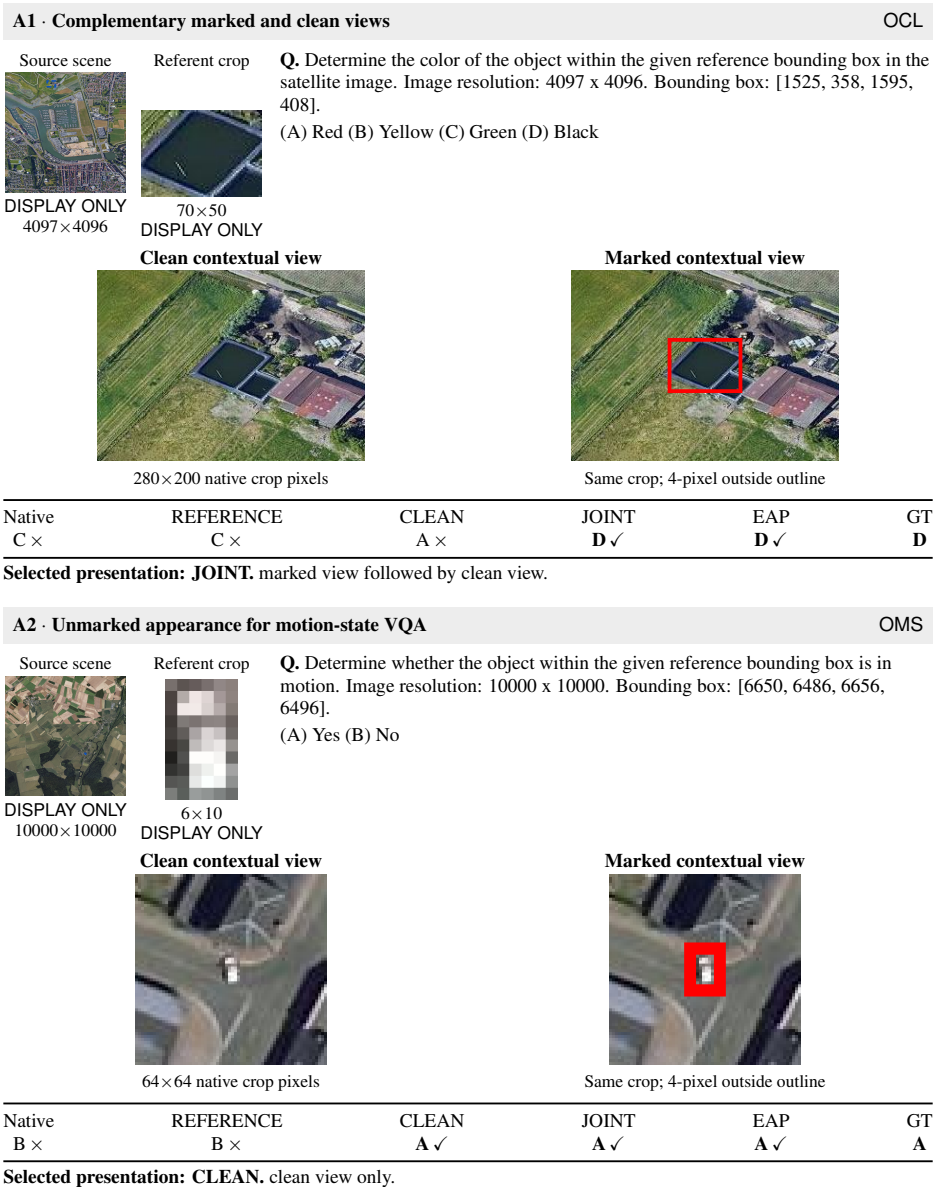


Figure S1: **Complementary marked and clean views.** A1: JOINT returns D (Black), while REFERENCE and CLEAN return C (Green) and A (Red). A2: CLEAN and JOINT return A (Yes); REFERENCE returns B (No). EAP selects JOINT for A1 and CLEAN for A2. JOINT receives the marked view first and the clean view second, showing the referenced region both with and without an outline. Displayed questions are the original inputs before EAP alignment. Source thumbnails, tight referent crops, and blue locators are display aids. ✓/× indicate agreement/disagreement with the evaluation answer (GT).

F.2 EXPLICIT REFERENCE WITHIN CONTEXTUAL CROPS

The category example A3 compares presentations of a contextual crop containing several regions. The color example A4 shows the same incorrect prediction across Native, the three presentation modes, and EAP.

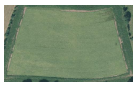
A3 · Explicit reference within a contextual crop OCC

Source scene



DISPLAY ONLY
10000×10000

Referent crop



484×286
DISPLAY ONLY

Q. Identify the object category within the given bounding box in the satellite or aerial image. Image resolution: 10000 x 10000. Bounding box: [2511, 977, 2995, 1263].

(A) The yellow open space (B) Orange Buildings (C) White buildings (D) Green Land

Clean contextual view



1936×1144 native crop pixels

Marked contextual view



Same crop; 4-pixel outside outline

Native	REFERENCE	CLEAN	JOINT	EAP	GT
C ×	**D** ✓	C ×	**D** ✓	**D** ✓	**D**

Selected presentation: REFERENCE. marked view only.

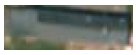
A4 · An error unchanged by presentation OCL

Source scene



DISPLAY ONLY
10000×10000

Referent crop



70×24
DISPLAY ONLY

Q. Determine the color of the object within the given reference bounding box in the satellite image. Image resolution: 10000 x 10000. Bounding box: [9572, 9631, 9642, 9655].

(A) Gray (B) Green (C) White (D) Black

Clean contextual view



280×96 native crop pixels

Marked contextual view



Same crop; 4-pixel outside outline

Native	REFERENCE	CLEAN	JOINT	EAP	GT
A ×	A ×	A ×	A ×	A ×	**D**

Selected presentation: JOINT. marked view followed by clean view.

Figure S2: **Predictions under alternative reference presentations.** A3: CLEAN returns C (White buildings), while REFERENCE and JOINT return the evaluation answer D (Green Land); EAP selects REFERENCE. A4: all three modes return A (Gray), whereas the evaluation answer is D (Black); EAP selects JOINT and returns A (Gray). Within each case, clean and marked views have identical spatial extent. The unfilled red outline is part of the marked model input; blue source-scene locators are display-only.

G LIMITATIONS AND FUTURE WORK

Spatial cues and contextual coverage. EAP uses a supplied spatial cue to identify the region to present. The XLRS perturbation study tests translation and scaling while keeping the answer for the original target unchanged (Appendix E). A cue that points to another object or excludes the target can still lead to an incorrect answer. The fixed context expansion can also miss nearby information needed for a question. Future work could adjust how much surrounding context is included in the crop based on the question.

Query-conditioned presentation. The current selector uses fixed query-text patterns and three presentation modes. These rules work for the evaluated query families but may not capture varied wording, multiple referents, or more complex questions. Future work could train a selector to choose REFERENCE, CLEAN, or JOINT from the meaning of the query, rather than matching fixed text patterns. Such a learned selector could support a wider range of questions while keeping the VLM backbone frozen. It would be a trained extension of the current training-free framework.

Finite evidence search. CES constructs a finite candidate pool through canonical search, scale refinement, and local translation. It can miss useful evidence outside these crops. Future work could vary the number and locations of crops according to the question, using fewer views for simple questions and more views when the relevant evidence is harder to locate. Combining crops from different parts of the scene may also help when an answer depends on the relationship between distant objects or regions.

Shared-model verification. Candidate answering, self-verification, and discriminative verification use the same frozen VLM. Verification scores guide evidence selection; they do not directly measure the probability that an answer is right. The same model can make similar mistakes when answering and verifying, so comparing two views does not guarantee that either answer is right. Future work could use a separate verifier or calibrate its scores to better reflect answer reliability.

Evaluation scope. The evaluation focuses on finite-choice, single-turn UHR-RS VQA on XLRS-Bench and RSHR-Bench. Extending GeoEPS to open-ended dialogue would require answering and verification beyond the current option-based interface. CES already handles questions without spatial cues, while EAP currently uses a supplied cue. Applying EAP to regions located automatically by CES could extend its presentation choices to questions that initially provide no spatial cue.

Inference cost. CES requires additional model calls to search for evidence and verify candidates. Future work could reduce the number of these calls while maintaining model's accuracy.