THE EARTH IN ONE GAZE: TRAINING-FREE ACTIVE FOCUS FOR UHR REMOTE SENSING UNDERSTANDING

Anonymous authors
Paper under double-blind review

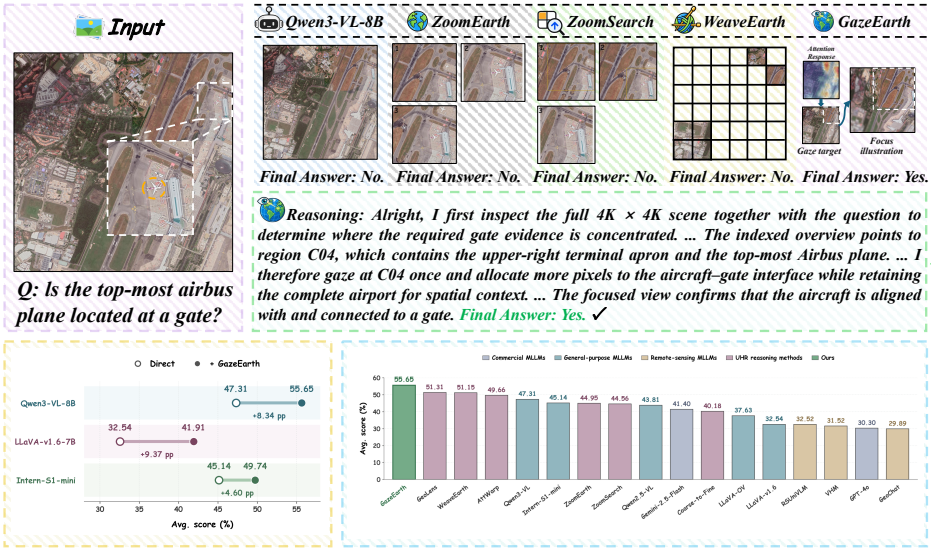

Figure 1: Overview of GazeEarth. (i) Different observation strategies can lead to different answers. (ii) GazeEarth combines question-guided region selection with foveated observation in a continuous full-scene view. (iii) The same strategy benefits different frozen backbones. (iv) Mean-score comparison with other methods on three remote-sensing benchmarks.

ABSTRACT

Multimodal large language models (MLLMs) must balance local detail against scene context when interpreting ultra-high-resolution (UHR) remote sensing (RS) imagery within a limited visual-input budget. Existing selection-based methods either prune tokens and select patches through relevance scoring, or crop actively through repeated inspection. Neither strategy directly redistributes pixels within a continuous full-scene view: the first retains selected tokens or patches, and the second re-encodes a crop detached from its surroundings. Our pilot study finds that a frozen MLLM already produces useful question-guided spatial requests, yet crop-based inspection of the selected regions does not consistently improve its answers. We therefore formulate UHR understanding as a question of where to spend a fixed pixel budget. Based on this, we introduce GazeEarth, a simple-yet-effective training-free framework that couples question-guided region selection with full-scene foveated observation. The MLLM selects evidence cells from an indexed overview; a deterministic, topology-preserving warp resamples the original image onto a fixed-size canvas, enlarging their shared neighborhood while compressing the periphery; the same frozen model answers from this focused view, using at most two MLLM calls and no external selector or iterative search. Across three UHR remote sensing benchmarks and four frozen backbones, GazeEarth improves benchmark-averaged accuracy by 4.6 to 9.4 percentage points over direct answering and 3.4 to 4.3 over overview answering, outperforming task-trained methods. Our analyses show that existing MLLMs can guide where to look in UHR images on their own, and that what they can infer from the selected evidence depends on how that evidence is presented.

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.

1 INTRODUCTION

Understanding ultra-high-resolution (UHR) remote sensing imagery requires balancing local fine-grained details against the broader scene context (Li et al., 2025b; Xin et al., 2026). For instance, determining whether a specific aircraft is parked at a terminal necessitates not only identifying the aircraft instance but also analyzing its spatial relationship to the surrounding airport infrastructure. However, a typical $10,000 \times 10,000$ scene contains roughly a hundred million pixels (Wang et al., 2025b; Luo et al., 2025), far exceeding the visual-token limits of current multimodal large language models (MLLMs). Standard uniform downsampling inevitably obscures critical small-scale objects (Guo et al., 2024; Ding et al., 2025; Li et al., 2026b; Liu et al., 2026c).

To circumvent this resolution bottleneck, recent methodologies have increasingly adopted active vision strategies. Methods such as targeted zooming (Wu & Xie, 2024; Shen et al., 2025; Zheng et al., 2026), learned crop-and-zoom policies (Liu et al., 2026b), or patch retrieval (Zhou et al., 2025) selectively allocate the limited pixel budget to specific regions of interest. Nevertheless, these approaches inherently suffer from contextual fragmentation: while extracting a high-resolution crop exposes local visual features, it detaches the target from its surrounding environment (Zhang et al., 2025b; 2026). In remote sensing, semantic meaning is heavily contingent on spatial relationships (Wang et al., 2024). Providing a global thumbnail alongside isolated crops requires the model to implicitly reconstruct spatial alignments, adding a cross-view association step during reasoning.

This raises a critical question: in UHR image understanding, is the primary bottleneck the localization of evidence, or its effective spatial presentation? Our pilot study (Section 2) reveals a counterintuitive insight. We demonstrate that off-the-shelf, frozen MLLMs can generate spatial requests aligned with question-specified regions, achieving higher alignment than dedicated external retrievers such as SigLIP2 (Tschannen et al., 2025). However, accurate localization does not guarantee successful task completion. When these question-selected regions are extracted as isolated crops and provided for reasoning, downstream performance remains inconsistent—even degrading by over 9 percentage points on MME-RealWorld-RS compared to the downsampled global overview. This indicates that identifying a relevant neighborhood is insufficient if inspecting it requires discarding its spatial surroundings.

Biological vision systems naturally resolve this via foveation: fixating on a region of interest to inspect fine-grained details while maintaining a compressed peripheral representation for global spatial awareness. (Schwartz, 1977; Recasens et al., 2018; Thavamani et al., 2021; 2023). Inspired by this mechanism, we propose GazeEarth, a simple-yet-effective training-free framework that reframes UHR understanding as a continuous spatial reallocation of the pixel budget rather than discrete image cropping. In GazeEarth, a frozen MLLM first selects an evidence region from an indexed overview. Subsequently, a deterministic, topology-preserving warp resamples the original image directly onto a fixed-size canvas, allocating more output pixels to the selected neighborhood while compressing the periphery. This ensures the focal region remains physically connected to its broader geographical context. The identical frozen model then performs reasoning over this single, foveated view. Without relying on iterative search algorithms, external selection modules, or parameter updates, this unified perception-action loop proves highly effective. Evaluated across three UHR remote sensing benchmarks utilizing four distinct frozen backbones, GazeEarth consistently improves over both direct and overview answering, delivering benchmark-averaged gains of 4.6–9.4 and 3.4–4.3 percentage points, respectively.

Our primary contributions are summarized as follows:

1. We distinguish coarse evidence localization from effective evidence presentation in UHR remote sensing, showing that a frozen MLLM provides useful spatial requests, but selected-region cropping does not consistently improve downstream answering.
2. We introduce GazeEarth, which couples a frozen MLLM's explicit region selection with deterministic, topology-preserving focus within a fixed-size answer-stage canvas, using at most two MLLM calls and no external selector or iterative search.
3. We demonstrate gains across three benchmarks and four frozen backbones, and characterize the roles of gaze source, evidence presentation, and selection depth through grounding, accuracy, and inference-cost analyses.

2 MOTIVATION: RETHINKING EVIDENCE PRESENTATION

To identify the true bottleneck in ultra-high-resolution (UHR) image reasoning, we design two preliminary probes. Specifically, we investigate whether the primary challenge lies in the model's inability to locate the evidence or in the way the extracted evidence is presented for downstream reasoning.

Finding 1: the answerer can already guide where to look. We compare the MLLM's spatial requests with question-derived reference cells and with requests from external observers and a question-only MLLM. As reported in Table 1, the image-conditioned MLLM achieves 77.74% benchmark-averaged alignment, compared with 42.82% for Question-only MLLM, 43.09% for SigLIP2, and 37.20% for RemoteCLIP. Its reported alignment exceeds question-only selection on all three benchmarks (Table 10). Figure 6 further shows that most off-target requests lie one grid step from a question-derived reference cell. These results characterize spatial-request alignment with question-specified regions. Further evaluation against dataset-provided target boxes on 537 LRS-GRO questions without explicit location cues yields a center-hit rate of 64.99%, compared with 52.33% for SigLIP2 and 37.99% for RemoteCLIP (Table 11). The next probe examines whether presenting the selected regions as crops improves answering.

Table 1: **Who selects the region.** Benchmark-averaged gaze grounding of different observers, together with the answer accuracy of their aligned and off-target requests.

Observer	Align. (%) ↑	Grid Dist. ↓	Acc.@ Aligned (%) ↑	Acc.@ Off-target (%) ↑
Random Focus	16.35	1.879	44.91	**42.93**
Question-only MLLM	42.82	1.450	–	–
SigLIP2	43.09	1.233	48.95	41.66
RemoteCLIP	37.20	1.343	48.80	41.47
Frozen MLLM	**77.74**	**0.334**	**50.49**	42.00

Finding 2: showing the model the crop does not reliably help. If the model can already generate a useful spatial request, does extracting and zooming into the selected region solve the UHR reasoning problem? To answer this, we conduct an evidence-presentation probe comparing reasoning on the downsampled full-scene view (Overview) against reasoning on the isolated high-resolution crop of the question-selected region (Gaze + Crop). As reported in Table 2, crop-based inspection severely degrades performance on contextually demanding tasks. For instance, on the MME-RealWorld-RS benchmark, feeding the MLLM the isolated high-resolution crop actually reduces the answering accuracy by 9.2 percentage points (from 49.01% to 39.81%) compared to simply reasoning over the blurry global overview. Furthermore, even when a global thumbnail is provided alongside the isolated crop (Think with Image), accuracy on MME-RealWorld-RS remains below Overview, failing to fully recover the performance drop.

Table 2: **How the selected evidence is presented.** Answer accuracy (%) of the frozen MLLM when the same question-selected region is shown in different ways.

Presentation	XLRS-Bench	LRS-GRO	MME-Real-World-RS	Avg.
Overview	45.13	59.99	**49.01**	51.38
Gaze + Crop	50.55	58.93	39.81	49.76
Think with image	**51.43**	**60.54**	42.24	**51.40**

Implication. Figure 2 summarizes gaze localization across the three benchmarks. Localization, however, determines where to inspect, not how the selected evidence is presented. While cropping magnifies local details, isolating a region can remove the surrounding landmarks needed for spatial reasoning. Accurate localization, therefore, needs an observation that exposes the selected evidence in context. The visual input should allocate more pixels to the target region while preserving its continuous spatial connection to the periphery. This motivates the topology-preserving foveated framework of GazeEarth.

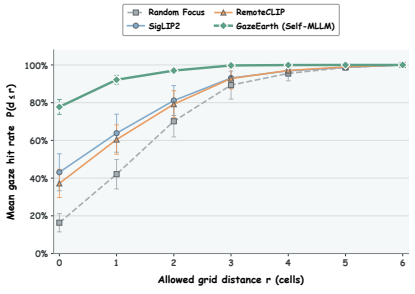

Figure 2: Gaze hit rates across three benchmarks (mean ± standard deviation).

3 GAZEEARTH

3.1 THE PERCEPTION-ACTION LOOP

We formulate UHR image understanding as a single-step perception-action loop driven entirely by a frozen MLLM, denoted as $F_\theta(\cdot)$. Given a UHR image $I \in \mathbb{R}^{H\times W\times 3}$ and a user query q, our

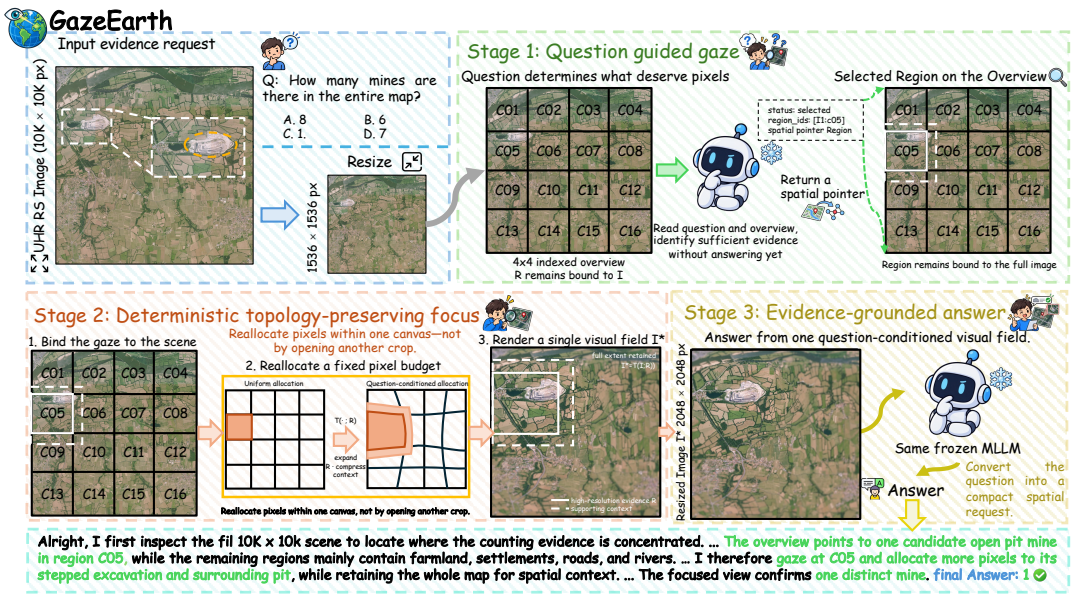

Figure 3: overview of GazeEarth. A frozen MLLM selects evidence regions from a question and an indexed overview. A deterministic, topology-preserving warp enlarges the selected neighborhood within a fixed-size, full-scene canvas. The same model then answers from this focused observation.

objective is to dynamically reallocate a fixed canvas pixel budget $B = h \times w$ (determined by the maximum answer-stage canvas size) without severing the global topology of the scene.

As illustrated in Figure 3, the GazeEarth framework execution unfolds in three consecutive phases:

- Question-Guided Gaze (Perception): The MLLM observes a severely downsampled overview $\tilde{I}_{idx}$ and the query q, acting as a zero-shot policy to output a spatial request R indicating where the critical visual evidence is concentrated.
- Deterministic Foveation (Action): A non-parametric, topology-preserving operator $\mathcal{T}_B$ executes this spatial request, rendering a foveated observation $I^* = \mathcal{T}_B(I; R)$. This operator magnifies the requested region R while compressing the global periphery, strictly adhering to the fixed budget B.
- Evidence-Grounded Answering (Perception): The identical frozen MLLM interprets the foveated continuous canvas I^* to generate the final semantic response $\hat{y}$.

3.2 Question Guided Gaze

The pilot study shows that a frozen MLLM can already produce useful question-guided spatial requests, yet selecting a region alone does not ensure successful answering. We therefore elicit a spatial request before answer generation and use it to guide the next observation. An indexed overview provides the spatial references through which the model specifies where to allocate more of the answer stage canvas. We overlay an indexed $m \times n$ grid, with $m = n = 4$, on an overview whose long side is capped at 1536 pixels. Each identifier maps to a source image region. For zero-based row r and column c, this region is

$$\mathcal{C}_{r,c} = \left[\frac{cW}{n}, \frac{(c+1)W}{n}\right) \times \left[\frac{rH}{m}, \frac{(r+1)H}{m}\right), \tag{1}$$

where $0 \le r < m$ and $0 \le c < n$, with coordinates ordered as (x, y) and raster boundaries rounded to pixels. The prompt requests the smallest set of at most two cells covering the evidence needed for q, without answering the question. This allows both regions in a comparison to be requested within one selection round. The valid identifiers are mapped to a set of grid indices $\mathcal{S}$. Parse constructs the spatial request R by padding the selected cells to include neighbouring context and merging overlapping regions:

$$R = \mathrm{Merge}(\{\mathrm{Pad}_\delta(\mathcal{C}_{r,c}) \mid (r,c) \in \mathcal{S}\}). \tag{2}$$

Here, Pad_δ extends each cell by a fraction δ of its width and height on each side and clips it to the source boundary; Merge replaces overlapping groups by their enclosing rectangles. A single gaze denotes one selection round, not one selected cell. An empty request from `global_only` or `brief_incomplete` retains the global view. A successfully resolved box in q directly defines a contextual rectangle and bypasses model-guided selection. Appendix B gives the parsing and coordinate rules. The resulting request R is passed to the focus operator, which enlarges the requested neighborhood while retaining the full scene.

3.3 Topology-Preserving Foveated Mapping

Identifying a relevant region is only useful if the next observation exposes its evidence in context. We therefore translate R into a continuous allocation of the answer-stage canvas: the selected neighborhood is enlarged while remaining spatially connected to the full scene.

Piecewise-linear foveated mapping. We determine the output canvas dimensions $h \times w$ from the source aspect ratio, with a 2048-pixel long-side cap. For nonempty R, the smallest enclosing rectangle defines a shared focus neighborhood, including the space between selected regions. Its boundaries are projected onto the canvas and rounded to pixels to determine the allocation, while the original source coordinates are retained for resampling. Along a canvas axis of length L, let $[a, b]$ denote the projected focus interval, with $0 \le a < b \le L$ and $\ell = b - a$. The enlargement factor $\alpha = 1.5$ and focus-fraction threshold $\beta = 0.55$ determine the target span:

$$\ell^\star = \min\{\max\{\ell, \mathrm{round}(\alpha\ell)\}, \max\{\ell, \mathrm{round}(\beta L)\}, L - \epsilon_- - \epsilon_+\}\,. \tag{3}$$

Here, $\epsilon_- = 1$ when $a > 0$ and $\epsilon_+ = 1$ when $b < L$, and each is zero otherwise. These margins reserve one output pixel for each nonempty peripheral interval. We centre the target interval $[a^\star, b^\star]$, where $b^\star = a^\star + \ell^\star$, on the projected interval when feasible and shift it to respect these margins. Appendix B gives the exact placement rule. The resulting canvas-coordinate mapping is

$$f(t) = \begin{cases} \dfrac{a^\star}{a}\, t, & 0 \le t < a, \\ a^\star + \dfrac{\ell^\star}{\ell}(t - a), & a \le t \le b, \\ b^\star + \dfrac{L - b^\star}{L - b}(t - b), & b < t \le L. \end{cases} \tag{4}$$

Empty outer segments are omitted. Applying this construction along both axes yields $\Phi_R(x, y) = (f_x(x), f_y(y))$, with $L = w$ horizontally and $L = h$ vertically.

Topology preservation. Each axis mapping is continuous and strictly increasing, with positive slopes on all nonempty segments and fixed canvas endpoints. Their product Φ_R is therefore a homeomorphism of the continuous canvas, preserving horizontal and vertical coordinate order without folds or cuts. The selected evidence and surrounding landmarks remain connected within one view as their scales change.

Pixel-budget allocation. The local area multiplier quantifies output-pixel allocation relative to the uniform canvas layout. Away from segment boundaries,

$$\rho_R(x, y) = \det D\Phi_R(x, y) = f_x'(x) f_y'(y). \tag{5}$$

On the focus rectangle, it equals $(\ell_x^\star/\ell_x)(\ell_y^\star/\ell_y) \ge 1$, where the subscripts identify the two axis spans. Whenever either span increases, the focus receives more output area. Because Φ_R covers the same canvas,

$$\int_0^h \int_0^w \rho_R(x, y)\, \mathrm{d}x\, \mathrm{d}y = hw = B. \tag{6}$$

Local expansion thus reduces the total area assigned to the periphery within the fixed budget. We render directly from the original image I, using the source boundaries retained in R and the target boundaries specified by Φ_R. Bilinear resampling proceeds first horizontally, retaining the original source height, and then vertically to produce the final $h \times w$ canvas: $I^\star = \mathrm{Render}_{\mathrm{bilinear}}(I, R, \Phi_R)$. An empty request returns an aspect-preserving bilinear resize of I to $h \times w$. The resulting focused view is passed to the answering stage.

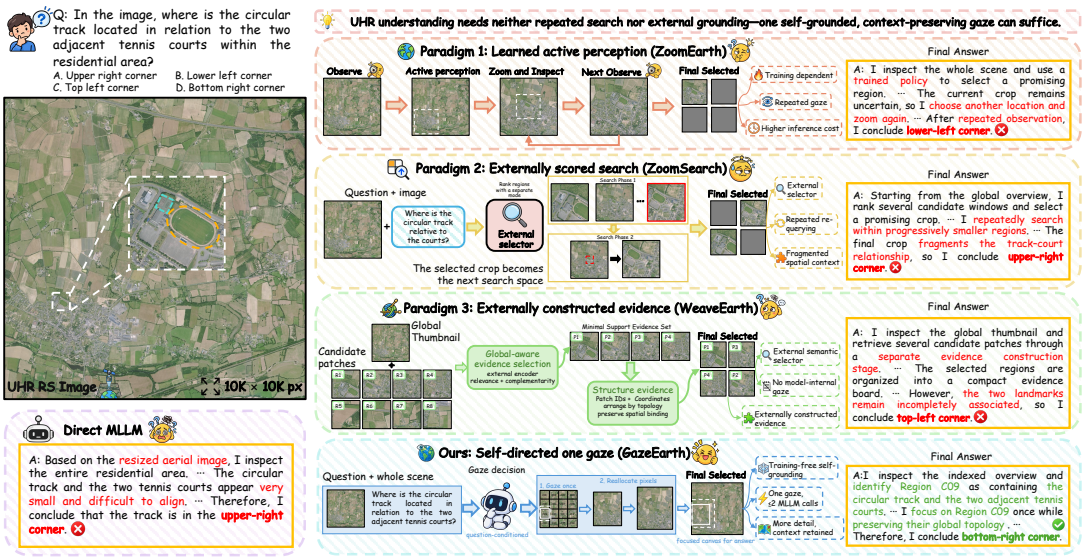

Figure 4: Observation strategies on a spatial-relation question. The example compares how different methods inspect a circular track and adjacent tennis courts. GazeEarth enlarges both landmarks within the full scene and correctly identifies their relative position.

3.4 Evidence-Grounded Answering

Following the topology-preserving spatial reallocation, the identical frozen MLLM performs the final reasoning step using the foveated canvas I^*. Because the foveation operator intrinsically preserves the global scene topology, the target evidence and its surrounding periphery are presented coherently within a single, continuous visual field. The MLLM is prompted with the original query q alongside the focused image I^* (annotated with its source label) to generate the final prediction. The instruction explicitly requires the model to ground its reasoning in the visible evidence and encapsulate the final output within `<answer>` tags, which is subsequently parsed by a lightweight extraction function to yield $\hat{y}$.

This final step closes a unified perception-action-perception sequence: the query directs the gaze, the gaze dynamically reconfigures the visual observation, and the foveated observation supports the final semantic reasoning. By design, this end-to-end framework requires at most two MLLM calls per question. Furthermore, this architectural separation establishes a common modular basis for evaluating the roles of the observer, gaze depth, and evidence composition, which we extensively analyse in our ablation studies.

4 Experiments

4.1 Experimental Setup

Benchmarks and Metrics. We evaluate GazeEarth on LRS-GRO (Liu et al., 2026b), MME-RealWorld-RS (Zhang et al., 2025c), and XLRS-Bench (Wang et al., 2025b), following the original evaluation protocols and metrics of each benchmark.

Backbones, Baselines, and Implementation. We use Qwen3-VL-8B as the primary backbone and assess cross-backbone generalization with LLaVA-v1.6-7B and Intern-S1-mini. We compare GazeEarth against commercial MLLMs, general-purpose MLLMs, remote-sensing MLLMs, and UHR visual reasoning methods. Unless otherwise specified, GazeEarth uses a 4×4 indexed overview with the longest side capped at 1,536 pixels and selects up to two regions in a single gaze round. We set the focus scale to 1.5 and cap the longest side of each answer-stage view at 2,048 pixels. The maximum generation length for the final answer is 1,024 tokens. The setting of Commercial MLLMs, prompts, and further implementation details are provided in the Appendix A.

Table 3: Main results on LRS-GRO and MME-RealWorld-RS. Scores (%) follow the respective benchmark evaluation protocols, with Avg. denoting the overall score for each benchmark. "–" in the result columns indicates an unreported score.

Method	Param.	TF	LRS-GRO (%) ↑				MME-RealWorld-RS (%) ↑			
			Global	Region	Object	Avg.	Color	Count	Position	Avg.
Commercial MLLMs										
GPT-4o (OpenAI, 2024)	–	✗	31.10	12.37	21.86	21.83	34.18	15.17	37.07	28.92
Gemini-2.5-Flash (Gemini Team, 2025)	–	✗	57.95	23.39	30.76	35.52	55.86	**33.28**	51.63	47.03
General-purpose MLLMs										
LLaVA-OV (Li et al., 2025a)	7B	✗	62.43	36.18	39.92	44.42	26.81	26.14	27.57	26.85
LLaVA-v1.6 (Liu et al., 2024)	7B	✗	75.52	35.67	45.24	50.23	24.51	18.03	27.49	23.39
Qwen2.5-VL (Bai et al., 2025b)	7B	✗	69.39	38.47	43.58	48.55	45.94	21.05	55.23	40.90
Qwen3-VL (Bai et al., 2025c)	8B	✗	76.25	42.03	54.77	56.90	48.07	21.35	56.41	42.11
Intern-S1-mini Bai et al. (2025a)	8B	✗	71.35	35.84	47.73	50.58	45.26	13.30	66.11	41.79
Remote-sensing MLLMs										
RSUniVLM (Liu & Lian, 2026)	0.5B	✗	63.02	34.10	37.55	42.83	28.97	19.00	28.33	25.48
GeoChat (Kuckreja et al., 2024)	7B	✗	62.43	36.79	40.72	44.99	23.11	15.66	25.06	21.32
VHM (Pang et al., 2025)	7B	✗	60.40	26.92	35.03	39.18	20.32	16.80	35.24	24.18
GeoLLaVA-8K (Wang et al., 2025a)	7B	✗	10.44	11.96	11.19	11.19	27.92	22.27	34.90	28.41
UHR visual reasoning methods										
Coarse-to-Fine (Luo et al., 2025)	7B	✗	71.35	35.32	36.52	44.57	44.70	31.00	49.72	41.89
GeoLens (Wang et al., 2026b)	7B	✗	**79.00**	48.90	57.60	60.67	53.23	19.74	63.64	45.75
ZoomEarth (Liu et al., 2026b)	3B	✗	67.45	42.31	51.96	53.39	49.69	31.66	50.14	43.93
ZoomSearch (Zhou et al., 2025)	8B	✓	67.14	46.78	52.04	54.41	41.67	17.86	54.97	38.34
RADAR (Liu et al., 2026d) †	8B	✓	–	–	–	–	50.52	20.47	58.15	43.23
WeaveEarth (Ma et al., 2026)	8B	✓	75.34	47.74	56.50	58.94	51.39	27.65	62.60	47.38
General Focus Method for MLLMs										
AttWarp (Dalal et al., 2026)	8B	✓	77.75	49.00	53.39	58.18	54.58	19.00	62.13	45.45
ViCrop (Zhang et al., 2025a)	8B	✓	77.45	**51.00**	53.82	58.80	51.87	17.54	62.13	44.06
Ours										
GazeEarth	8B	✓	77.96	50.39	**58.04**	**61.00**	**63.67**	21.78	**71.12**	**52.43**

Table 4: Performance (%) on XLRS-Bench across eight categories: Counting (Cnt), Scene Classification (SC), Object Spatial Relationship (OSR), Object Properties (OP), Planning (Plan), Anomaly Reasoning (AR), Complex Reasoning (CR), and Spatiotemporal Reasoning (SR).

Method	Param.	TF	XLRS-Bench (%) ↑								
			Cnt	SC	OSR	OP	Plan	AR	CR	SR	Avg.
Commercial MLLMs											
GPT-4o (OpenAI, 2024)	–	✗	32.50	47.67	35.20	37.47	36.00	72.00	58.50	31.67	40.16
Gemini-2.5-Flash (Gemini Team, 2025)	–	✗	38.75	51.67	30.40	39.10	42.00	78.00	65.50	23.33	41.66
General-purpose MLLMs											
LLaVA-OV (Li et al., 2025a)	7B	✗	31.25	35.67	25.20	45.42	24.00	76.00	59.50	43.33	41.62
LLaVA-v1.6 (Liu et al., 2024)	7B	✗	24.38	21.33	23.80	21.27	40.00	35.00	34.50	33.33	24.00
Qwen2.5-VL (Bai et al., 2025b)	7B	✗	30.63	40.33	30.80	43.31	32.00	67.00	62.50	43.33	41.98
Qwen3-VL (Bai et al., 2025c)	8B	✗	32.50	41.67	31.40	43.92	34.00	70.00	63.50	46.33	42.92
CogVLM2 (Hong et al., 2024)	8B	✗	36.88	46.33	36.20	36.87	34.00	69.00	60.50	-	40.20
Intern-S1-mini (Bai et al., 2025a)	8B	✗	35.63	58.00	32.80	38.80	51.00	69.00	66.00	**58.33**	43.05
Remote-sensing MLLMs											
RSUniVLM (Liu & Lian, 2026)	0.5B	✗	21.88	24.67	34.40	30.18	20.00	28.00	28.00	25.00	29.25
GeoChat (Kuckreja et al., 2024)	7B	✗	22.50	12.67	24.20	24.28	10.00	33.00	32.00	–	23.35
VHM (Pang et al., 2025)	7B	✗	32.50	33.33	31.40	26.69	33.00	58.00	49.50	31.67	31.20
GeoLLaVA-8K (Wang et al., 2025a)	7B	✗	35.00	62.00	35.80	37.71	**65.00**	68.00	63.50	51.67	43.44
UHR visual reasoning methods											
Coarse-to-Fine (Luo et al., 2025)	7B	✗	37.50	36.33	25.40	30.66	27.00	70.00	58.50	51.67	34.09
GeoLens (Wang et al., 2026b)	7B	✗	36.25	62.33	34.40	45.96	58.00	66.00	65.00	48.33	47.50
ZoomEarth (Liu et al., 2026b)	3B	✗	35.00	44.67	31.40	35.60	22.00	66.00	56.50	28.30	37.53
ZoomSearch (Zhou et al., 2025)	8B	✓	37.50	59.33	25.80	41.99	34.00	69.00	68.00	43.33	43.15
WeaveEarth (Ma et al., 2026)	8B	✓	37.50	50.33	36.40	47.41	36.00	73.00	67.50	46.33	47.14
General Focus Method for MLLMs											
AttWarp (Dalal et al., 2026)	8B	✓	33.75	58.67	31.00	44.94	39.00	74.00	62.50	46.67	45.36
ViCrop (Zhang et al., 2025a)	8B	✓	31.88	59.00	29.40	44.82	37.00	74.00	63.00	45.00	44.90
Ours											
GazeEarth	8B	✓	**44.38**	**66.67**	**37.80**	**53.07**	49.00	76.00	**74.50**	55.00	**53.51**

4.2 Main Results

Overall performance. As shown in Tables 3 and 4, GazeEarth achieves the best overall accuracy among the compared methods on all three benchmarks, reaching 61.00% on LRS-GRO, 52.43% on MME-RealWorld-RS, and 53.51% on XLRS-Bench. Compared with direct inference using the same frozen Qwen3-VL-8B backbone, these results correspond to absolute gains of 4.10, 10.32, and 10.59 percentage points, respectively. The gains are not limited to aggregate performance: relative to Overview, GazeEarth improves Object accuracy on LRS-GRO by 1.82 percentage points and Color accuracy on MME-RealWorld-RS by 8.13 points. The additional changes in LRS-GRO Region and MME-RealWorld-RS Position are +0.17 and -1.27 points, respectively. Appendix Figure 7 reports category-level gains over Direct, while Tables 15 – 17 distinguish these gains from the additional changes over Overview.

Case study. Figure 4 illustrates the change in observation on a spatial-relation question. GazeEarth selects the correct region and enlarges the circular track and adjacent tennis courts within the full scene. Both landmarks become easier to inspect while their relative position is retained, supporting the correct answer, "bottom-right corner."

Table 5: GazeEarth across four frozen backbones. Gain denotes the percentage-point improvement over the corresponding Direct baseline. Mean is the unweighted average of the three benchmark scores.

Backbone	Method	LRS-GRO	MME-RS	XLRS-Bench	Mean
Qwen3-VL-8B	*Direct*	56.90	42.11	42.92	47.31
	Overview	59.99	49.01	45.13	51.38
	+ GazeEarth	**61.00**	**52.43**	**53.51**	**55.65**
	Gain	+4.10	+10.32	+10.59	+8.34
LLaVA-v1.6-7B	*Direct*	50.23	23.39	24.00	32.54
	Overview	50.99	30.16	34.35	38.50
	+ GazeEarth	**54.36**	**33.90**	**37.47**	**41.91**
	Gain	+4.13	+10.51	+13.47	+9.37
Intern-S1-mini	*Direct*	50.58	41.79	43.05	45.14
	Overview	52.08	43.10	42.69	45.96
	+ GazeEarth	**55.95**	**46.44**	**46.82**	**49.74**
	Gain	+5.37	+4.65	+3.77	+4.60
GPT-4o	*Direct*	21.83	28.92	40.16	30.30
	Overview	23.20	32.27	41.03	32.17
	+ GazeEarth	**26.79**	**36.18**	**44.29**	**35.75**
	Gain	+4.96	+7.26	+4.13	+5.45

Cross-backbone performance. The gains extend beyond the primary Qwen backbone. With each backbone generating its own spatial requests and answering from the resulting focused views, GazeEarth improves all three benchmark scores for Qwen3-VL-8B, LLaVA-v1.6-7B, Intern-S1-mini and GPT-4o (Table 5). Their benchmark-averaged gains over Direct are 8.34, 9.37, 4.60, and 5.45 percentage points, respectively, with additional gains of 4.27, 3.41, 3.78, and 3.58 points over Overview. The same observation procedure thus benefits all four evaluated backbones without updating their parameters.

Table 6: Observer substitution with Qwen3-VL-8B as the fixed answerer.

Observer	XLRS-Bench (%) ↑	LRS-GRO (%) ↑	MME-Real-World-RS (%) ↑	Avg.	Δ vs. Random
Random	45.42	58.97	45.64	50.01	–
SigLIP2	46.10	59.54	47.49	51.04	+1.03
RemoteCLIP	45.94	60.10	47.08	51.04	+1.03
GazeEarth	**53.51**	**61.00**	**52.43**	**55.65**	**+5.64**

Question-guided allocation. We next vary who selects the regions while keeping Qwen3-VL-8B as the answerer and retaining the same focused-view answering pipeline (Table 6). Random Focus achieves a benchmark-averaged score of 50.01%; SigLIP2 and RemoteCLIP each reach 51.04%; GazeEarth reaches 55.65%. Self-directed gaze improves over both external observers on every benchmark, with a mean advantage of 4.61 percentage points. Together with the localization probe, this shows that the frozen MLLM's spatial requests are useful both for locating evidence and for guiding observations that support answering. Appendix Figure 8 reports sensitivity to grid size, region budget, focus scale, and selector resolution.

Allocation matters more than input volume.

The three answering conditions use the same frozen backbone but differ in both the amount and spatial allocation of visual input. Direct achieves 47.31% at 7.34 seconds per question. Overview improves accuracy to 51.38% while reducing runtime to 4.02 seconds. GazeEarth preserves the same canvas and pixel budget (Eq. 6) but reallocates pixels according to the question, reaching 55.65% in 5.07 seconds (Tables 7 and 22). Thus, more visual input is not necessarily better: a smaller, well-structured input is both faster and more accurate, and question-guided reallocation yields a further gain. Of the 8.34-point improvement over Direct, 4.07 points come from canvas normalization and the remaining 4.27 from directed focus.

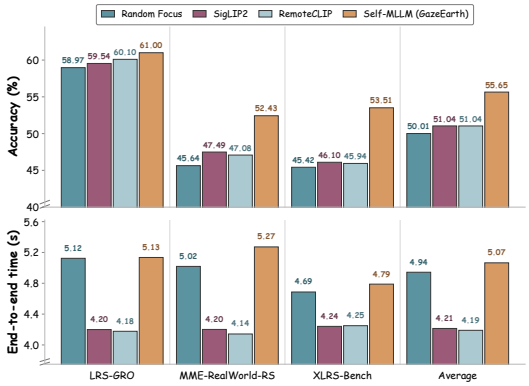

Figure 5: Accuracy and end-to-end inference time under observer substitution.

4.3 Ablation Study

Evidence presentation. Table 7 compares full-scene focus with an unfocused overview and crop-based evidence presentation. GazeEarth achieves a benchmark-averaged score of 55.65%, compared

Table 7: Ablation study on the three benchmarks. Gaze denotes the number of region-selection rounds; "Think with Image" indicates that both global and cropped views are provided to the model.

Variant	Gaze	Q-guided gaze	Detail	Full scene	Continuous Binding	Performance (%) ↑			
						XLRS-Bench	LRS-GRO	MME-RealWorld-RS	Avg.
Direct	0	✗	✗	✓	–	42.92	56.90	42.11	47.31
Overview	0	✗	✗	✓	–	45.13	59.99	49.01	51.38
Gaze + Crop	1	✓	✓	✗	✗	50.55	58.93	39.81	49.76
Think with Image	1	✓	✓	✓	✗	51.43	60.54	42.24	51.40
Linked-Global + Crop	1	✓	✓	✓	✓	52.56	60.25	43.39	52.07
Two-Gaze	2	✓	✓	✓	✓	46.27	60.12	52.33	52.91
Three-Gaze	3	✓	✓	✓	✓	46.36	60.04	51.77	52.72
GazeEarth	1	✓	✓	✓	✓	**53.51**	**61.00**	**52.43**	**55.65**

with 51.38% for Overview, 49.76% for Gaze + Crop, and 51.40% for Think with Image (a global overview and a separate crop without explicit spatial linkage). The difference is largest on MME-RealWorld-RS: continuous full-scene focus reaches 52.43%, versus 39.81% and 42.24% for the two crop-based variants. On XLRS-Bench, GazeEarth reaches 53.51%, compared with 50.55% for Gaze + Crop, 51.43% for Think with Image, and 52.56% for Linked Global + Crop. Appendix D additionally evaluates AttWarp warping. GazeEarth also achieves the highest three-benchmark mean among the presentation variants evaluated on all three benchmarks. These results show that region selection alone does not determine answering performance: how the selected evidence is presented also matters. The comparison with Direct measures the overall gain of the GazeEarth pipeline, whereas the comparison with Overview measures its additional benefit over full-scene answering.

5 Related Work

Remote Sensing Multimodal Large Language Models. Remote sensing MLLMs leverage descriptive and geographic supervision (Hu et al., 2025; Muhtar et al., 2024), grounded dialogue (Kuckreja et al., 2024), multisensor alignment (Zhang et al., 2024), and geospatial reasoning (Liu et al., 2026a). For UHR imagery, GeoLLaVA-8K retains informative tokens (Wang et al., 2025a), ZoomSearch searches and reassembles patches (Zhou et al., 2025), and WeaveEarth integrates local evidence, global context, and spatial metadata (Ma et al., 2026). These passive token purification strategies rely heavily on how patches are selected and struggle to preserve spatial context.

Active Vision in General Domains and Remote Sensing. Active vision uses language-guided search (Wu & Xie, 2024), cropping (Zhang et al., 2023), and learned inspection (Su et al., 2025; Lee et al., 2026; Lin et al., 2026). Internal attention also provides localization cues (Zhang et al., 2025a; Shi et al., 2026). Training-free methods collect spatially arranged crops (Wang et al., 2025c) or hierarchical evidence (Li et al., 2026a). In remote sensing, learned zooming (Liu et al., 2026b) and tool use (Wang et al., 2026b;a) require post-training, while repeated cropping adds inference rounds and cross-view integration. Nonuniform sampling reallocates resolution (Recasens et al., 2018; Thavamani et al., 2021); AttWarp uses cross-modal attention to guide context-preserving image warping (Dalal et al., 2026). GazeEarth instead elicits an explicit spatial request through the model's text-generation interface. This request specifies the focus neighborhood for a deterministic piecewise-linear mapping that samples the original UHR image into a fixed-size full-scene canvas.

6 Conclusion

In this work, we view UHR remote sensing understanding as a problem of active foveated representations. Our analysis shows that frozen MLLMs can already localize relevant evidence, but existing crop-based thinking-with-image methods do not consistently improve answering, which highlights the importance of how selected evidence is presented. Based on this insight, we introduced GazeEarth, a simple-yet-effective training-free framework that converts explicit spatial requests into topology-preserving, full-scene focused observations. Across three benchmarks and four frozen backbones, GazeEarth consistently improves over direct and overview answering, with benchmark-averaged gains of 4.6–9.4 and 3.4–4.3 percentage points, respectively. These results suggest that effective observation construction, rather than repeated visual search alone, is a promising direction for improving MLLM reasoning over UHR imagery.

AI USE STATEMENT

In this work, we used generative AI tools for language editing, improving clarity and readability, and refining the wording of the manuscript. We have not used generative AI tools for generating research data, conducting data analysis, or drawing scientific conclusions, and other disclosure categories are not applicable to this work. Additionally, we used generative AI tools to suggest alternative phrasing and improve the overall presentation of the text. We have reviewed all AI-assisted work carefully. All AI-generated or AI-assisted text was checked and revised by the authors to ensure accuracy, consistency with the underlying research, and appropriate attribution of relevant sources. No AI-generated content was accepted without author review. We take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This study was carried out in accordance with the principles outlined in the ICLR Code of Ethics. Throughout the research process, we sought to maintain high standards of research integrity, transparency, and responsible scientific practice.

- We have made every effort to ensure that the results reported in this paper faithfully reflect our observations and experiments. No data, analyses, or experimental outcomes have been fabricated, altered, or selectively presented in a misleading manner. We also provide sufficient methodological and experimental details to facilitate independent reproduction of our results.
- The datasets employed in this work were accessed and used in compliance with their respective licenses, usage policies, and applicable terms.
- We have appropriately recognized the contributions of previous studies and external resources that informed this research. Relevant prior work has been cited throughout the paper, and all individuals listed as co-authors have contributed meaningfully to the study and approved the final submission.
- We have also examined the possible societal and practical consequences of the proposed methods. Although the primary objective of this research is to contribute constructively to the field, we recognize that certain techniques may have dual-use potential. We therefore emphasize the importance of careful, responsible, and context-aware deployment in practical applications.

REPRODUCIBILITY STATEMENT

To support the reproducibility and independent validation of our findings, we report the experimental configurations, key hyperparameter settings, and relevant implementation details with sufficient specificity. Whenever feasible, we also release the corresponding source code and supplementary resources to enable result verification, replication, and subsequent research built upon our work.

REFERENCES

Lei Bai et al. Intern-S1: A scientific multimodal foundation model, 2025a. URL https://arxiv.org/abs/2508.15763.

Shuai Bai et al. Qwen2.5-VL technical report, 2025b. URL https://arxiv.org/abs/2502.13923.

Shuai Bai et al. Qwen3-VL technical report, 2025c. URL https://arxiv.org/abs/2511.21631.

Dwip Dalal, Gautam Vashishtha, Utkarsh Mishra, Jeonghwan Kim, Madhav Kanda, Hyeonjeong Ha, Svetlana Lazebnik, Heng Ji, and Unnat Jain. Constructive distortion: Improving MLLMs with attention-guided image warping. In *International Conference on Learning Representations*, pp. 23639–23672, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/file/288b63aa98084366c4536ba0574a0f22-Paper-Conference.pdf.

Xinpeng Ding, Jianhua Han, Hang Xu, Wei Zhang, and Xiaomeng Li. HiLM-D: Enhancing MLLMs with multi-scale high-resolution details for autonomous driving. *International Journal of Computer Vision*, 133(8):5379–5395, 2025. doi: 10.1007/s11263-025-02433-3. URL https://doi.org/10.1007/s11263-025-02433-3.

Gemini Team. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities, 2025. URL https://arxiv.org/abs/2507.06261.

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, Cham, 2024. Springer Nature Switzerland. doi: 10.1007/978-3-031-73010-8_23. URL https://doi.org/10.1007/978-3-031-73010-8_23.

Wenyi Hong et al. CogVLM2: Visual language models for image and video understanding, 2024. URL https://arxiv.org/abs/2408.16500.

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.

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. doi: 10.1109/CVPR52733.2024.02629.

Jewon Lee, Wooksu Shin, Seungmin Yang, Ki-Ung Song, DongUk Lim, Jaeyeon Kim, Tae-Ho Kim, and Bo-Kyeong Kim. ERGO: Efficient high-resolution visual understanding for vision-language models. In *International Conference on Learning Representations*, pp. 8750–8769, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/0f0c4f3d83c58df58380af3b0729354c-Abstract-Conference.html.

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. URL https://openreview.net/forum?id=zKv8qULV6n.

Kaiyu Li, Zixuan Jiang, Xiangyong Cao, Jiayu Wang, Yuchen Xiao, Deyu Meng, and Zhi Wang. DescribeEarth: Describe anything for remote sensing images, 2025b. URL https://arxiv.org/abs/2509.25654.

Yangfu Li, Hongjian Zhan, Jiawei Chen, Yuning Gong, Qi Liu, and Yue Lu. DeepScan: A training-free framework for visually grounded reasoning in large vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 19155–19164, June 2026a. URL https://openaccess.thecvf.com/content/CVPR2026/papers/Li_DeepScan_A_Training-Free_Framework_for_Visually_Grounded_Reasoning_in_Large_CVPR_2026_paper.pdf.

Yanwei Li, Yuechen Zhang, Chengyao Wang, Zhisheng Zhong, Yixin Chen, Ruihang Chu, Shaoteng Liu, and Jiaya Jia. Mini-Gemini: Mining the potential of multi-modality vision language models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 48(3):3530–3543, 2026b. doi: 10.1109/TPAMI.2025.3637265. URL https://doi.org/10.1109/TPAMI.2025.3637265.

Zichuan Lin, Yicheng Liu, Yang Yang, Lvfang Tao, and Deheng Ye. AdaptVision: Efficient vision-language models via adaptive visual acquisition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11923–11932, June 2026. URL https://openaccess.thecvf.com/content/CVPR2026/html/Lin_AdaptVision_Efficient_Vision-Language_Models_via_Adaptive_Visual_Acquisition_CVPR_2026_paper.html.

Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. LLaVA-NeXT: Improved reasoning, OCR, and world knowledge, January 2024. URL https://llava-vl.github.io/blog/2024-01-30-llava-next/.

Jiaqi Liu, Lang Sun, Ronghao Fu, and Bo Yang. Towards faithful reasoning in remote sensing: A perceptually-grounded GeoSpatial chain-of-thought for vision-language models. In *International Conference on Learning Representations*, pp. 6535–6563, 2026a. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/0b8067e0598da978cf89f5fc12c0acc2-Abstract-Conference.html.

Ruixun Liu, Bowen Fu, Jiayi Song, Kaiyu Li, Wanchen Li, Lanxuan Xue, Hui Qiao, Weizhan Zhang, Deyu Meng, and Xiangyong Cao. ZoomEarth: Active perception for ultra-high-resolution geospatial vision-language tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 34877–34888, 2026b. URL https://openaccess.thecvf.com/content/CVPR2026/html/Liu_ZoomEarth_Active_Perception_for_Ultra-High-Resolution_Geospatial_Vision-Language_Tasks_CVPR_2026_paper.html.

Wenzhuo Liu, Weijie Yin, Fei Zhu, Shijie Ma, Haiyang Guo, Yi Chen, Xiao-Hui Li, Xiao Liang, Chao Feng, and Cheng-lin Liu. One patch doesn't fit all: Adaptive patching for native-resolution multimodal large language models. In *International Conference on Learning Representations*, pp. 27594–27608, 2026c. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/2e936c33952e9370f863c0b0410c09c1-Abstract-Conference.html.

Xu Liu and Zhouhui Lian. RSUniVLM: A unified vision-language model for remote sensing via granularity-oriented MoE. *Pattern Recognition*, 179:113717, 2026. doi: 10.1016/j.patcog.2026.113717. URL https://doi.org/10.1016/j.patcog.2026.113717.

Yi Liu, Jing Zhang, Di Wang, Xiaoyu Tian, Haonan Guo, and Bo Du. Seeing clearly without training: Mitigating hallucinations in multimodal LLMs for remote sensing, 2026d. URL https://arxiv.org/abs/2603.02754.

Junwei Luo, Yingying Zhang, Xue Yang, Kang Wu, Qi Zhu, Lei Liang, Jingdong Chen, and Yansheng Li. When large vision-language model meets large remote sensing imagery: Coarse-to-fine text-guided token pruning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9206–9217, 2025. doi: 10.1109/ICCV51701.2025.00860.

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:2607.10120*, 2026. URL https://arxiv.org/abs/2607.10120.

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*, pp. 440–457. Springer Nature Switzerland, 2024. doi: 10.1007/978-3-031-72904-1_26. URL https://doi.org/10.1007/978-3-031-72904-1_26.

OpenAI. GPT-4o System Card. https://openai.com/index/gpt-4o-system-card/, August 2024. Accessed: 2026-09-17.

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. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39 (6):6381–6388, 2025. doi: 10.1609/aaai.v39i6.32683. URL https://doi.org/10.1609/aaai.v39i6.32683.

Adrià Recasens, Petr Kellnhofer, Simon Stent, Wojciech Matusik, and Antonio Torralba. Learning to zoom: A saliency-based sampling layer for neural networks. In *Computer Vision – ECCV 2018*, pp. 52–67. Springer, 2018. doi: 10.1007/978-3-030-01240-3_4. URL https://doi.org/10.1007/978-3-030-01240-3_4.

Eric L. Schwartz. Spatial mapping in the primate sensory projection: Analytic structure and relevance to perception. *Biological Cybernetics*, 25(4):181–194, 1977. doi: 10.1007/BF01885636.

Haozhan Shen, Kangjia Zhao, Tiancheng Zhao, Ruochen Xu, Zilun Zhang, Mingwei Zhu, and Jianwei Yin. ZoomEye: Enhancing multimodal LLMs with human-like zooming capabilities through tree-based image exploration. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 6602–6618. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.emnlp-main.335.

Yuheng Shi, Xiaohuan Pei, Minjing Dong, and Chang Xu. Catching the details: Self-distilled RoI predictors for fine-grained MLLM perception. In *International Conference on Learning Representations*, pp. 136133–136155, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/dc6319dde4fb182b22fb902da9418566-Abstract-Conference.html.

Alex Su, Haozhe Wang, Weiming Ren, Fangzhen Lin, and Wenhu Chen. Pixel reasoner: Incentivizing pixel space reasoning via curiosity-driven reinforcement learning. In *Advances in Neural Information Processing Systems*, volume 38, pp. 8222–8251. Curran Associates, Inc., 2025. doi: 10.52202/085713-0280. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/0c38f54740062529aa4117a04b583f3c-Abstract-Conference.html.

Chittesh Thavamani, Mengtian Li, Nicolas Cebron, and Deva Ramanan. FOVEA: Foveated image magnification for autonomous navigation. *arXiv preprint arXiv:2108.12102*, 2021. URL https://arxiv.org/abs/2108.12102.

Chittesh Thavamani, Mengtian Li, Francesco Ferroni, and Deva Ramanan. Learning to zoom and unzoom. *arXiv preprint arXiv:2303.15390*, 2023. URL https://arxiv.org/abs/2303.15390.

Michael Tschannen, Alexey Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim Alabdulmohsin, Nikhil Parthasarathy, Talfan Evans, Lucas Beyer, Ye Xia, Basil Mustafa, Olivier Hénaff, Jeremiah Harmsen, Andreas Steiner, and Xiaohua Zhai. SigLIP 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features, 2025. URL https://arxiv.org/abs/2502.14786.

Fengxiang Wang, Mingshuo Chen, Yueying Li, Di Wang, Haotian Wang, Zonghao Guo, Zefan Wang, Boqi Shan, Long Lan, Yulin Wang, Hongzhen Wang, Wenjing Yang, Bo Du, and Jing Zhang. GeoLLaVA-8K: Scaling remote-sensing multimodal large language models to 8K resolution. In *Advances in Neural Information Processing Systems*, volume 38, pp. 159185–159218. Curran Associates, Inc., 2025a. doi: 10.52202/085713-5317. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/e95ef5c6ed4096c609e0b8b47ffaeb9b-Abstract-Conference.html.

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 (CVPR)*, 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, Yuhao Zhou, Di Wang, Yifan Zhang, Haoyu Wang, Haiyan Zhao, Hongda Sun, Long Lan, Jun Song, Yulin Wang, Jing Zhang, Wenlong Zhang, and Bo Du. Text before vision: Staged knowledge injection matters for agentic rlvr in ultra-high-resolution remote sensing understanding, 2026a. URL https://arxiv.org/abs/2602.14225.

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, 2026b. URL https://arxiv.org/abs/2607.25993.

Junjue Wang, Zhuo Zheng, Zihang Chen, Ailong Ma, and Yanfei Zhong. EarthVQA: Towards queryable earth via relational reasoning-based remote sensing visual question answering. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(6):5481–5489, 2024. doi: 10.1609/aaai.v38i6.28357. URL https://ojs.aaai.org/index.php/AAAI/article/view/28357.

Wenbin Wang, Yongcheng Jing, Liang Ding, Yingjie Wang, Li Shen, Yong Luo, Bo Du, and Dacheng Tao. Retrieval-augmented perception: High-resolution image perception meets visual RAG. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 63290–63307. PMLR, 2025c. URL https://proceedings.mlr.press/v267/wang25at.html.

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 (CVPR)*, 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.

Zepeng Xin, Kaiyu Li, Luodi Chen, Wanchen Li, Yuchen Xiao, Hui Qiao, Weizhan Zhang, Deyu Meng, and Xiangyong Cao. SegEarth-R2: Towards comprehensive language-guided segmentation for remote sensing images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13199–13210, June 2026. URL https://openaccess.thecvf.com/content/CVPR2026/html/Xin_SegEarth-R2_Towards_Comprehensive_Language-guided_Segmentation_for_Remote_Sensing_Images_CVPR_2026_paper.html.

Jiarui Zhang, Mahyar Khayatkhoei, Prateek Chhikara, and Filip Ilievski. Towards perceiving small visual details in zero-shot visual question answering with multimodal LLMs, 2023. URL https://arxiv.org/abs/2310.16033.

Jiarui Zhang, Mahyar Khayatkhoei, Prateek Chhikara, and Filip Ilievski. MLLMs know where to look: Training-free perception of small visual details with multimodal LLMs. In *International Conference on Learning Representations*, pp. 68194–68213, 2025a. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/aaa0ac4253da75faf9b0dc0dda062612-Abstract-Conference.html.

Renshan Zhang, Rui Shao, Gongwei Chen, Miao Zhang, Kaiwen Zhou, Weili Guan, and Liqiang Nie. FALCON: Resolving visual redundancy and fragmentation in high-resolution multimodal large language models via visual registers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 23530–23540, October 2025b. URL https://openaccess.thecvf.com/content/ICCV2025/html/Zhang_FALCON_Resolving_Visual_Redundancy_and_Fragmentation_in_High-resolution_Multimodal_Large_ICCV_2025_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.

Yi-Fan Zhang, Huanyu Zhang, Haochen Tian, Chaoyou Fu, Shuangqing Zhang, Junfei Wu, Feng Li, Kun Wang, Qingsong Wen, Zhang Zhang, Liang Wang, Rong Jin, and Tieniu Tan. MME-RealWorld: Could your multimodal LLM challenge high-resolution real-world scenarios that are difficult for humans? In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2025c. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/df29d63af05cb91d705cf06ba5945b9d-Abstract-Conference.html.

YiFan Zhang, Qingsong Wen, Chaoyou Fu, Kun Wang, Xue Wang, Zhang Zhang, Liang Wang, and Rong Jin. Beyond LLaVA-HD: Diving into high-resolution multimodal large language models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 48(5):5493–5504, 2026. doi: 10.1109/TPAMI.2026.3650761. URL https://doi.org/10.1109/TPAMI.2026.3650761.

Ziwei Zheng, Michael Yang, Jack Hong, Chenxiao Zhao, Guohai Xu, Le Yang, Chao Shen, and Xing Yu. DeepEyes: Incentivizing "thinking with images" via reinforcement learning. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/cdb347d7516d52a9280ea9d6708911f6-Abstract-Conference.html.

Yunqi Zhou, Chengjie Jiang, Chun Yuan, and Jing Li. Look where it matters: Training-free ultra-HR remote sensing VQA via adaptive zoom search. *arXiv preprint arXiv:2511.20460*, 2025. URL https://arxiv.org/abs/2511.20460.

A Experimental and Implementation Details

This section details the prompts, image preprocessing and decoding settings, benchmark evaluation protocols, and hardware and timing configurations used in our experiments. The supplementary analyses separate three questions: whether the model produces useful spatial requests (Appendix C), how these requests should be rendered for answering (Appendix D), and what accuracy–cost trade-offs follow from the resulting observation pipeline (Appendices E–G). Throughout, we report where each comparison is controlled and where several factors change together, so that end-to-end gains are not attributed to a single component without a corresponding control.

A.1 Prompts

GazeEarth uses two stage-specific prompts for the same model: one elicits a spatial request from the indexed overview, and the other elicits an answer from the resulting full-scene observation. The request determines the visual input of the answering call, which receives the question and the rendered view as a fresh interaction.

Question-guided gaze. The selection prompt combines the original question with the indexed source overview and requests a compact JSON response. The placeholder {`question`} contains the question text, and {`source_ids`} lists the source identifiers, such as `I1` or `I1, I2`.

```
You are a visual relevance selector for a remote-sensing question, not
↪ an answerer.

Question: {question}
Indexed source overview(s): {source_ids}. Each source uses the same 4x4
↪ neutral grid.

Choose the smallest visual brief that would let another expert answer:
- global_only: the overview already contains sufficient visual
↪ evidence;
- selected: magnify one or two cells needed for a local detail, local
↪ relation, or two-region comparison;
- brief_incomplete: the question requires evidence scattered over more
↪ than two regions, exhaustive counting/distribution, or cannot be
↪ represented by this brief.

Return JSON only. Never answer the question, describe objects, or
↪ provide reasoning.
Valid forms:
{"status":"global_only","region_ids":[]}
{"status":"selected","region_ids":["I1:C07"]}
{"status":"selected","region_ids":["I1:C04","I1:C12"]}
{"status":"brief_incomplete","region_ids":[]}
```

Evidence-grounded answering. The answering prompt accompanies one rendered full-scene view per source image and requests a reasoning explanation followed by the final answer enclosed in `<answer>` tags. For multiple-choice questions, {`answer_question`} contains the question followed by `Options:` and the letter-labelled choices in their original order. For open-ended questions, it contains the question alone. GazeEarth and the Overview condition share this answering template.

```
The visual input contains one full-scene view for each remote-sensing
↪ source. Any labels are neutral correspondence cues only.
Answer the visual question using only visible evidence. Explain your
↪ reasoning before giving the final answer.
Question: {answer_question}
Enclose only the final answer in <answer>...</answer> tags.
```

Table 8: Backbone interfaces and inference settings. Local backbones follow their checkpoint-default visual processing, whereas GPT-4o uses provider-managed visual preprocessing. Scene-size caps apply before source-label headers and backbone-specific or provider-side preprocessing. Generation limits count new tokens.

Setting	Qwen3-VL 8B	LLaVA-v1.6 Mistral-7B	Intern-S1 mini	Commercial MLLMs API
Execution interface	Local	Local	Local	OpenRouter
Vision patch size	16×16	14×14	14×14	Not disclosed
Vision input unit	Variable-size image	336×336 tile	448×448 tile	PNG/JPEG data URL
Resolution handling	Dynamic resolution	AnyRes tiling	Dynamic tiling	Provider-managed
Default processor limit	16,777,216 pixels	4 local tiles	12 local tiles	Not disclosed
Additional global view	–	1	1 when tiled	Not disclosed
API image-detail mode	–	–	–	`high`
Gaze overview long-side cap	1536 pixels			
Answer canvas long-side cap	2048 pixels			
Selection grid	4×4			
Maximum selected cells	2			
Decoding	Greedy	Greedy	Greedy	Temperature 0
Gaze `max_new_tokens`	64			
Answer `max_new_tokens`	1024			

Qwen3-VL applies 2×2 spatial patch merging. LLaVA's AnyRes candidate resolutions are 336×672, 672×336, 672×672, 1008×336, and 336×1008. Intern-S1-mini adds a global thumbnail when more than one local tile is produced. GPT-4o is accessed through the OpenRouter Chat Completions API using `openai/gpt-4o-2024-11-20` with `detail=high` and temperature 0. Visual inputs are transmitted as base64-encoded PNG images, with JPEG quality 95 used only when an encoded image exceeds 29 MB. Provider-side patching and the effective encoder resolution are not exposed by the API.

A.2 Backbone Interfaces and Inference Protocol

Visual interfaces. We use Qwen3-VL-8B, LLaVA-v1.6-Mistral-7B, and Intern-S1-mini through their checkpoint-specific image processors and chat templates. Table 8 summarizes their visual interfaces and generation settings. Each source image is converted to RGB. For question-guided gaze, we construct an indexed overview with a long-side cap of 1536 pixels and a 4×4 grid. The resulting overview and question form the selection input. For evidence-grounded answering, the spatial request guides resampling from the original image onto an aspect-preserving canvas with a long-side cap of 2048 pixels. The rendered view carries a neutral source identifier without the selection grid. These size limits apply to the scene content before source-label headers and backbone-specific preprocessing.

Model calls and decoding. The same backbone performs region selection and answering, with its parameters fixed throughout inference. Each stage uses a separate interaction containing its visual input and stage-specific prompt. The answering call receives the rendered observation and the original question, without the selector's conversation history. The selector's effect is therefore mediated through the constructed visual observation, rather than through forwarded reasoning text or conversation history. Local backbones use greedy decoding with sampling disabled, so temperature and nucleus sampling do not affect token selection. closed source models use the API configuration listed in Table 8, including temperature 0. All backbones use output limits of 64 new tokens for question-guided gaze and 1024 for evidence-grounded answering.

Hyperparameter configuration. The default configuration uses a 4×4 indexed grid, a maximum of $K = 2$ selected cells, a focus scale of $\alpha = 1.5$, and a selector overview with a 1536-pixel long-side cap. The grid provides 16 spatial addresses for expressing a compact observation request. Allowing two cells accommodates both local inspection and comparisons between two regions. The focus scale controls the enlargement of the selected neighbourhood relative to its uniformly resized extent, while the selector resolution controls the image size used to generate the request. The same configuration is used across the three benchmarks. The sensitivity experiments evaluate variations around this configuration.

Table 9: Benchmark composition and evaluation scope. Counts denote question–answer instances in the evaluation sets used in this work.

Benchmark	Questions	Answer format	Reporting categories
LRS-GRO	9,734	Open-ended	Global, Region, Object
MME-RealWorld-RS	3,738	Multiple-choice	Color, Count, Position
XLRS-Bench	3,080	Multiple-choice	Cnt, SC, OSR, OP, Plan, AR, CR, SR

A.3 Benchmarks and Evaluation Protocols

Benchmark composition. We evaluate on LRS-GRO, MME-RealWorld-RS, and XLRS-Bench, covering open-ended and multiple-choice remote-sensing question answering. Table 9 summarizes the evaluation sets and reporting categories. LRS-GRO contains questions at global, region, and object levels. MME-RealWorld-RS evaluates color, counting, and position recognition. XLRS-Bench covers eight categories spanning visual perception and reasoning. The sample counts refer to question–answer instances. The localization probe uses separately defined eligible cohorts rather than the complete answer-evaluation sets.

Scoring and aggregation. For LRS-GRO, we follow the benchmark's answer-matching rule: a prediction is correct if it matches the reference after lowercasing or satisfies the WordNet path-similarity threshold of 0.8. For MME-RealWorld-RS and XLRS-Bench, predictions are matched to the reference multiple-choice answers. Category scores are computed over the questions belonging to each category. Within each benchmark, Avg. denotes the overall question-level accuracy, equivalently the category accuracies weighted by their sample counts. Across benchmarks, Avg. or Mean denotes the unweighted arithmetic mean of the three overall scores. Improvements are reported in absolute percentage points. Cross-benchmark averages are computed only for methods with results on all three benchmarks.

Baseline provenance. We evaluate GazeEarth, its controlled variants, and the corresponding Direct baselines using the selected backbone checkpoints. The main comparison tables distinguish results evaluated in this work from externally reported scores: † indicates that the experimental results for this model are taken from the original paper. When a method combines results from different sources, the provenance is indicated at the individual-score level. A dash denotes an unreported result. All metrics not reported in the original paper were reproduced using the official code and parameter configurations.

A.4 Hardware and Timing Protocol

Hardware. Timing experiments are conducted on a single NVIDIA A800 GPU with a batch size of one question. The model is loaded once and reused across samples. For GazeEarth, region selection and answering are executed sequentially using the same model instance.

Timing protocol. We measure per-question wall-clock latency after the source images have been loaded and converted to RGB, ending when the extracted answer is available. The timed interval includes overview construction, question-guided gaze when invoked, focused-view rendering, backbone-specific preprocessing, and answer generation. Model initialization, source-image I/O, benchmark scoring, and prediction-file writing are excluded. Each evaluated question contributes one timing observation, with no dedicated warm-up phase or repeated timing trials. We report the arithmetic mean of per-question latencies within each benchmark and the unweighted mean of the three benchmark-specific means across benchmarks. Selection and answering times are also recorded separately; the total additionally includes visual construction and inter-stage processing.

B Foveated Mapping Details

This section details spatial-request parsing, target-interval placement, and rendering from the original source images for the topology-preserving foveated mapping operator. Topology preservation is a

property of the continuous coordinate mapping: the mapping is monotonic along each axis and has no folds, so it preserves connectivity and horizontal and vertical order.

B.1 Parsing and Coordinate Rules

The indexed grid links each selected cell to a rectangle in the original image through Eq. 1. Cells are numbered in row-major order, and identifiers such as `I1:C07` specify the source image and cell. After normalisation and deduplication, a valid `selected` response contains one or two identifiers. The statuses `global_only` and `brief_incomplete` require an empty identifier list. Invalid responses produce an empty request.

Cell boundaries are rounded to source-image pixels. Each selected cell is padded by a fraction δ of its width and height on each side, with $\delta = 0.25$ by default. The padded boundaries are rounded and clipped to the image extent, and overlapping rectangles from the same source are merged. The enclosing rectangle of the resulting request defines the shared focus neighborhood. Its original-image coordinates are retained for rendering.

A labeled bounding box supplied in the question is mapped directly to source coordinates. Normalized coordinates are scaled by W and H; absolute coordinates are interpreted at the declared image resolution. The mapped boundaries are rounded and clipped. For multiple images, an explicit source identifier determines which image the box refers to.

The resolved box is expanded into a contextual square. Let d denote its longer side and $M = \max(W, H)$. The default side length is $4d$, clamped between $u = \max\{d, \mathrm{round}(0.04M)\}$ and $v = \max\{u, \mathrm{round}(0.30M)\}$, then capped by $\min(W, H)$. The square is centred on the box and shifted inside the image when necessary. It supplies the spatial request directly; otherwise, the indexed-overview selector is used.

B.2 Placement Rule for the Target Interval

The target interval preserves the source interval's centre whenever the available peripheral space permits. For the span $\ell^\star$ defined in Eq. 3, its endpoints are

$$a^\star = \min\left\{L - \epsilon_+ - \ell^\star,\ \max\left\{\epsilon_-,\ \mathrm{round}\left(\frac{a+b-\ell^\star}{2}\right)\right\}\right\}, \qquad b^\star = a^\star + \ell^\star. \tag{7}$$

This places the enlarged interval within the canvas while reserving one output pixel for each nonempty peripheral segment. A focus interval touching a canvas edge remains attached to that edge. The construction is applied independently along both axes, with empty outer segments omitted.

B.3 Rendering and Multi-Source Handling

Rendering pairs the retained original-image boundaries with the target boundaries specified by Φ_R. The projection onto the $h \times w$ canvas determines the target spans and positions, while the original image supplies the pixels. When a projected boundary rounds to a canvas edge, its corresponding source boundary is set to the original-image edge.

The separable mapping is rendered in two passes. First, the original image is divided at the focus boundaries along the horizontal axis into up to three vertical strips. Each strip is bilinearly resampled to its assigned target width while retaining the original height H. Their ordered concatenation forms an $H \times w$ intermediate image. The second pass uses the original row boundaries to resample the corresponding horizontal strips to their target heights, producing the final $h \times w$ view. Both passes use Pillow's bilinear resampling with explicit source bounds.

For multiple source images, requests are grouped by source identifier and rendered independently. An empty request returns an aspect-preserving bilinear resize under the same long-side cap; images already within the cap retain their dimensions. The resulting views are passed together to the answerer in source order. Each view carries the same neutral, 36-pixel-high source header used by the Overview condition. The header is appended above the scene canvas, whose pixel budget is $B = hw$.

C Pilot Study Details

This section details the two pilot probes: spatial-request alignment and evidence presentation. We additionally validate target localization against dataset-provided bounding boxes on LRS-GRO.

C.1 Setup

For the localisation probe, we overlay a 4×4 indexed grid on a full-scene overview with a 1536-pixel long-side cap. The frozen MLLM receives the overview and question and returns cell identifiers without answering. We compare its requests with those of SigLIP2, RemoteCLIP, Random Focus, and Question-only MLLM. The latter receives the question and grid definition without an image.

For the spatial-cue alignment analysis, the reference region is constructed from explicit spatial expressions in the question. Ordered rules map corners, quadrants, directional bands, and central regions to reference cells; the first matching rule determines the reference region. In this analysis, each observer is evaluated on questions with both a matched reference and an extractable selected cell. Questions without a matched reference and requests without an extractable selected cell are excluded from these statistics. The resulting evaluation measures agreement with the question-specified region.

For a nonempty selected-cell set and reference-cell set, d is the minimum Manhattan distance between their cells. Thus, $d = 0$ means that at least one selected cell belongs to the reference region. We report the alignment rate $P(d = 0)$, mean grid distance, and tolerance curves $P(d \leq r)$. Table 10 reports per-benchmark alignment and its unweighted three-benchmark mean. Figure 6 reports the tolerance curves and eligible sample counts for the plotted observers.

For the evidence-presentation probe, we compare *Overview*, *Gaze + Crop*, and *Think with Image* using the same answerer across XLRS-Bench, LRS-GRO, and MME-RealWorld-RS. *Overview* uses a uniformly resized full-scene view. *Gaze + Crop* presents the question-selected crop alone, whereas *Think with Image* presents that crop together with a separate overview without explicit spatial linkage.

C.2 Spatial-Request Alignment and Target Localization

Table 10 expands the aggregate alignment comparison in Table 1. The image-conditioned MLLM achieves alignment rates of 73.89%, 77.60%, and 81.74% on XLRS-Bench, LRS-GRO, and MME-RealWorld-RS, respectively. Its three-benchmark mean is 77.74%, compared with 43.09% for SigLIP2, 37.20% for RemoteCLIP, and 16.35% for Random Focus.

Question-only MLLM achieves 41.50%, 44.56%, and 42.40% on the three benchmarks, with a mean of 42.82%. The reported alignment rates of the image-conditioned MLLM are higher by 32.39, 33.04, and 39.34 percentage points, respectively, yielding a mean difference of 34.92 points. Image-conditioned selection therefore exhibits stronger agreement with the question-specified regions in this probe.

Table 10: Spatial-request alignment with question-derived reference cells across benchmarks. We report $P(d = 0)$ on the eligible examples for each observer. Avg. is the unweighted mean of the three benchmark scores; higher is better.

Observer	XLRS-Bench (%) ↑	LRS-GRO (%) ↑	MME-RealWorld-RS (%) ↑	Avg. (%) ↑
Random Focus	14.65	21.83	12.56	16.35
Question-only MLLM	41.50	44.56	42.40	42.82
SigLIP2	39.49	54.18	35.59	43.09
RemoteCLIP	33.44	45.82	32.33	37.20
Frozen MLLM	**73.89**	**77.60**	**81.74**	**77.74**

Figure 6 shows how alignment changes with the allowed grid distance. At $r = 1$, the MLLM's hit rates reach 89.49% on XLRS-Bench, 93.65% on LRS-GRO, and 93.33% on MME-RealWorld-RS; at $r = 2$, all three exceed 96%. The increase from $r = 0$ to $r = 1$ accounts for more than half of its off-target requests on each benchmark. Its selected cells are thus concentrated within and immediately around the question-specified regions.

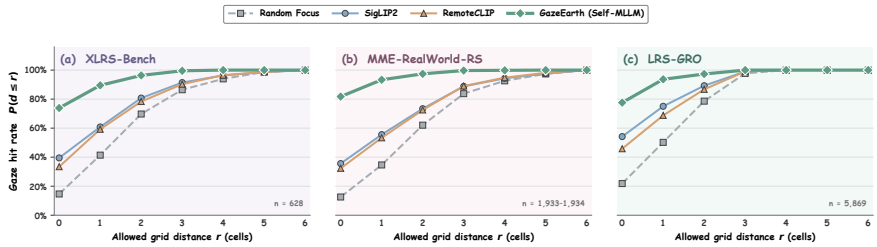

Figure 6: Spatial-request alignment within an allowed grid distance r on three benchmarks. Here, d is the minimum Manhattan distance between selected cells and question-derived reference cells. Curves show $P(d \leq r)$ on the eligible examples for each observer; sample counts refer to these examples.

Table 11: Target localization and oracle-gaze answering on 537 LRS-GRO questions from 159 images without explicit location cues. The three observers share the same per-question request status and cell budget for localization. Higher is better for all percentage metrics.

Observer	Center hit (%) ↑	Box coverage (%) ↑	Answer accuracy (%) ↑
SigLIP2	52.33	47.56	64.43
RemoteCLIP	37.99	34.33	62.94
Self-MLLM	**64.99**	**55.65**	66.29
Oracle Gaze	–	–	**70.20**

C.3 Localization against independent target annotations.

To complement the spatial-cue analysis, we evaluate the recorded selections against dataset-provided target boxes on LRS-GRO. The evaluation subset contains 537 object-attribute questions from 159 images without explicit location cues or coordinate references. It covers object category, color, material, shape, and state or activity. Questions are selected by their content and annotation validity, rather than model predictions or correctness. We verify each question–target correspondence and the mapping of annotation coordinates to the original image. The aspect-preserving, 1024-pixel long-side annotation coordinates are converted to source-image coordinates by multiplying both axes by $\max(W, H)/1024$. Reference boxes are used only for evaluation. The external observers retain GazeEarth's request status and per-question cell budget: 493 questions have one selected cell, 29 have two, and 15 have no selected cell. Center hit measures whether the selected cells contain the annotated target center. Box coverage measures the fraction of the target-box area covered by the union of selected cells, averaged over questions. Both metrics use the original 4×4 grid cells before padding or region merging. All 537 questions remain in the denominator, with empty or invalid requests receiving zero.

Table 11 shows that GazeEarth achieves a center-hit rate of 64.99%, exceeding SigLIP2 and RemoteCLIP by 12.66 and 27.00 percentage points, respectively. Its mean target-box coverage reaches 55.65%, compared with 47.56% and 34.33%. These results extend the spatial-cue analysis to independently annotated targets: the answerer's own requests also localize question-relevant objects when their positions are not explicitly stated. We further evaluate answering with oracle gaze on the same annotated subset. Dataset-provided target boxes replace predicted regions in constructing the full-scene focused view, which is then passed to the frozen Qwen3-VL-8B answerer using the CoT2 prompt. Oracle-bbox Gaze achieves 70.20% answering accuracy, complementing the localization results with an assessment of answering under annotation-guided focus.

C.4 Aligned versus Off-target Requests

For the question-derived reference-cell analysis, Acc.@Aligned measures answer accuracy on requests with $d = 0$, and Acc.@Off-target measures accuracy on requests with $d > 0$, separately for each observer. Spatial Acc. is the answer accuracy over that observer's evaluated spatial subset. Table 1 reports the unweighted three-benchmark mean of each metric. Tables 12, 13, and 14 provide the corresponding per-benchmark results.

For the MLLM's own requests, aligned examples achieve 50.49% benchmark-averaged answer accuracy, compared with 42.00% for off-target examples. The same association holds within each benchmark: 44.40% versus 37.20% on XLRS-Bench, 57.65% versus 50.00% on LRS-GRO, and

49.43% versus 38.81% on MME-RealWorld-RS. Aligned requests are more often answered correctly, but alignment alone does not ensure a correct answer. These conditional accuracies reflect the combined outcome of evidence perception and answer reasoning. They motivate examining how the selected region is subsequently presented to the answerer.

Table 12: Spatial-request alignment and conditional answering accuracy on XLRS-Bench.

Observer	Align. (%) ↑	Grid Dist. ↓	Acc.@ Aligned (%) ↑	Acc.@ Off-target (%) ↑	Spatial Acc. (%) ↑
Random Focus	14.65	1.951	39.13	**38.99**	39.01
SigLIP2	39.49	1.326	40.32	37.37	38.54
RemoteCLIP	33.44	1.435	41.90	35.89	37.90
Frozen MLLM	**73.89**	**0.408**	**44.40**	37.20	**42.52**

Table 13: Spatial-request alignment and conditional answering accuracy on MME-RealWorld-RS.

Observer	Align. (%) ↑	Grid Dist. ↓	Acc.@ Aligned (%) ↑	Acc.@ Off-target (%) ↑	Spatial Acc. (%) ↑
Random Focus	12.56	2.170	41.98	36.72	37.38
SigLIP2	35.59	1.548	49.27	36.22	40.87
RemoteCLIP	32.33	1.599	48.64	36.16	40.20
Frozen MLLM	**81.74**	**0.280**	**49.43**	**38.81**	**47.49**

C.5 Evidence Presentation

Table 2 compares answering from the full-scene overview, the selected crop alone, and the crop accompanied by a separate overview. This probe examines how the answerer uses question-selected evidence under different presentations.

The effects differ across benchmarks. On XLRS-Bench, presenting the selected crop raises accuracy from 45.13% to 50.55%, and adding the overview increases it to 51.43%. On MME-RealWorld-RS, isolated cropping reduces accuracy from 49.01% to 39.81%; adding the overview recovers part of this decrease, reaching 42.24%. On LRS-GRO, the corresponding scores are 59.99%, 58.93%, and 60.54%.

Across the three benchmarks, *Overview*, *Gaze + Crop*, and *Unlinked Global + Crop* achieve mean accuracies of 51.38%, 49.76%, and 51.40%, respectively. Presenting the requested region as a crop therefore helps on XLRS-Bench but hurts on MME-RealWorld-RS, and does not improve the three-benchmark mean. With the selection procedure held fixed, answering can become better or worse depending on how the requested region is rendered. Together, the two probes separate spatial-request alignment from the usefulness of the resulting observation, and motivate treating the conversion of a request into the next observation as a design problem in its own right. GazeEarth is one such conversion: it enlarges the selected neighbourhood within a continuous full-scene view.

Table 14: Spatial-request alignment and conditional answering accuracy on LRS-GRO.

Observer	Align. (%) ↑	Grid Dist. ↓	Acc.@ Aligned (%) ↑	Acc.@ Off-target (%) ↑	Spatial Acc. (%) ↑
Random Focus	21.83	1.516	53.63	**53.07**	53.19
SigLIP2	54.18	0.825	57.26	51.39	54.57
RemoteCLIP	45.82	0.996	55.86	52.36	53.96
Frozen MLLM	**77.60**	**0.314**	**57.65**	50.00	**55.93**

D Extended Quantitative Results

This section expands the main quantitative results through category-level comparisons, observer substitution, gaze-depth analysis, and the comparison between random focus and the full-scene overview.

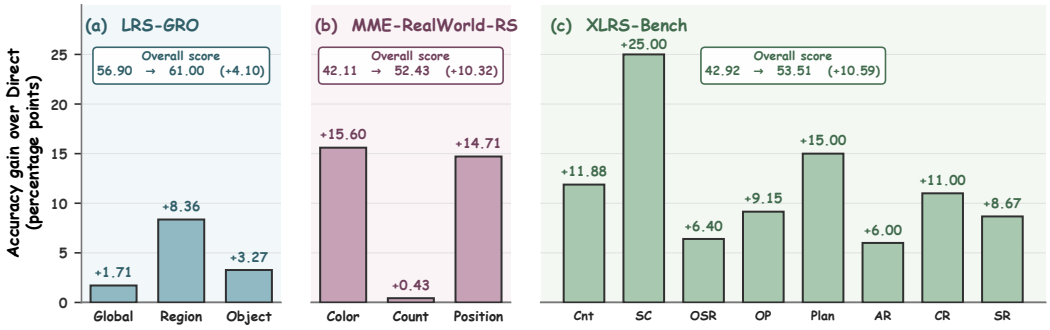

Figure 7: Category-wise accuracy gains of GazeEarth over Direct using the same Qwen3-VL-8B backbone on (a) LRS-GRO, (b) MME-RealWorld-RS, and (c) XLRS-Bench. Bars show absolute improvements in percentage points. Insets report the overall benchmark scores (Direct → GazeEarth) and their differences.

D.1 Category-Level Breakdown

Figure 7 reports category-wise gains over Direct using Qwen3-VL-8B. GazeEarth improves all reported categories across the three benchmarks. On XLRS-Bench, the largest gains occur in Scene Classification and Planning, at 25.00 and 15.00 percentage points, followed by Counting and Complex Reasoning at 11.88 and 11.00 points. Counting improvements differ across benchmarks: MME-RealWorld-RS increases from 21.35% to 21.78%, whereas XLRS-Bench increases from 32.50% to 44.38%.

Table 15: Category-level controlled analysis on XLRS-Bench.

Variant	Gaze	Q-guided gaze	Detail	Full scene	Binding	XLRS-Bench (%) ↑								
						Cnt	SC	OSR	OP	Plan	AR	CR	SR	Avg.
Visual Evidence Composition														
Direct	0	✗	✗	✓	–	32.50	41.67	31.40	43.92	34.00	70.00	63.50	46.33	42.92
Overview	0	✗	✗	✓	–	35.00	60.67	28.80	44.52	42.00	71.00	64.50	45.00	45.13
Random Focus	1	✗	✓	✓	✓implicit	35.00	61.33	30.60	44.10	39.00	**78.00**	65.50	43.33	45.42
Gaze + Crop	1	✓	✓	✗	✗	41.88	66.33	29.20	52.29	41.00	72.00	69.50	41.67	50.55
Think with Image	1	✓	✓	✓	✗	36.88	63.67	29.20	55.66	40.00	71.00	64.50	40.00	51.43
Linked Global + Crop	1	✓	✓	✓	✓explicit	33.75	62.33	30.00	**57.83**	39.00	73.00	66.00	40.00	52.56
AttWarp	1	✓	✓	✓	✓implicit	33.75	58.67	31.00	44.94	39.00	74.00	62.50	46.67	45.36
Gaze Depth and Iterative Baselines														
Two-Gaze	2	✓	✓	✓	✓implicit	38.12	64.33	31.20	44.82	44.00	72.00	65.50	40.00	46.27
Three-Gaze	3	✓	✓	✓	✓implicit	40.62	65.00	31.00	44.16	45.00	74.00	68.00	41.67	46.36
GazeEarth	1	✓	✓	✓	✓implicit	**44.38**	**66.67**	**37.80**	53.07	**49.00**	76.00	**74.50**	**55.00**	**53.51**

Tables 15, 16, and 17 provide category-level comparisons of the evidence-presentation variants. *Overview* presents the uniformly resized full scene, while *Gaze + Crop* presents only the question-selected region. *Think with Image* (*Unlinked Global + Crop*) provides both views without explicit spatial linkage; *Linked Global + Crop* additionally identifies the correspondence between the crop and its source region in the overview. We additionally evaluate AttWarp with Qwen3-VL-8B as the backbone. AttWarp guides image warping through cross-modal attention, whereas GazeEarth uses an explicit textual spatial request to guide deterministic piecewise-linear full-scene resampling. GazeEarth instead uses the model's explicit spatial request to guide piecewise-linear full-scene warping. On XLRS-Bench, GazeEarth achieves the highest overall accuracy and leads on Counting, Scene Classification, OSR, Planning, Complex Reasoning, and SR among the listed variants. Linked Global + Crop leads on Object Properties, while Random Focus leads on Anomaly Reasoning. On LRS-GRO, GazeEarth achieves the highest Global, Object, and overall scores, while Linked-Global + Crop leads on Region questions. The comparison with Overview separates these gains from those obtained by changing the full-scene input. On LRS-GRO, Region accuracy increases from 42.03% with Direct to 50.22% with Overview and 50.39% with GazeEarth. Object accuracy increases from 54.77% to 56.22% and 58.04%, respectively. Thus, Region shows the largest gain over Direct, whereas Object shows the largest additional gain over Overview. On MME-RealWorld-RS, Gaze + Crop improves Color accuracy over Overview from 55.54% to 62.47%, but reduces Position accuracy from 72.39% to 34.45%. The effect of evidence presentation, therefore, varies across question categories. GazeEarth focuses on details while preserving global topology, thereby offering an overall advantage.

Table 16: Category-level controlled analysis on LRS-GRO.

Variant	Gaze	Q-guided gaze	Detail	Full scene	Binding	LRS-GRO (%) ↑			
						Global	Region	Object	Avg.
Visual Evidence Composition									
Direct	0	✗	✗	✓	–	76.25	42.03	54.77	56.90
Overview	0	✗	✗	✓	–	77.88	50.22	56.22	59.99
Random Focus	1	✗	✓	✓	✓implicit	77.84	48.78	54.95	58.97
Gaze + Crop	1	✓	✓	✗	✗	74.66	47.39	56.95	58.93
Think with Image	1	✓	✓	✓	✗	76.80	51.09	57.38	60.54
Linked Global + Crop	1	✓	✓	✓	✓explicit	75.34	**51.52**	57.30	60.25
AttWarp	1	✓	✓	✓	✓implicit	77.75	49.00	53.39	58.18
Gaze Depth and Iterative Baselines									
Two-Gaze	2	✓	✓	✓	✓implicit	76.93	50.00	57.01	60.12
Three-Gaze	3	✓	✓	✓	✓implicit	76.68	49.17	57.34	60.04
GazeEarth	**1**	✓	✓	✓	✓implicit	**77.96**	50.39	**58.04**	**61.00**

Table 17: Category-level controlled analysis on MME-RealWorld-RS.

Variant	Gaze	Q-guided gaze	Detail	Full scene	Binding	MME-RealWorld-RS (%) ↑			
						Color	Count	Position	Avg.
Visual Evidence Composition									
Direct	0	✗	✗	✓	–	48.07	21.35	56.41	42.11
Overview	0	✗	✗	✓	–	55.54	18.35	**72.39**	49.01
Random Focus	1	✗	✓	✓	✓implicit	52.43	16.80	66.98	45.64
Gaze + Crop	1	✓	✓	✗	✗	62.47	22.10	34.45	39.81
Unlinked Global + Crop	1	✓	✓	✓	✗	63.35	20.64	42.24	42.24
Linked Global + Crop	1	✓	✓	✓	✓explicit	62.79	19.74	47.10	43.39
AttWarp	1	✓	✓	✓	✓implicit	54.58	19.00	62.13	45.45
Gaze Depth and Iterative Baselines									
GazeEarth	**1**	✓	✓	✓	✓implicit	63.67	21.78	71.12	**52.43**
Two-Gaze	2	✓	✓	✓	✓implicit	**65.74**	23.98	66.59	52.33
Three-Gaze	3	✓	✓	✓	✓implicit	64.78	**24.71**	65.16	51.77

D.2 Observer Substitution per Benchmark

We vary the observer while keeping Qwen3-VL-8B as the answerer and using the same focused-view construction and answering pipeline. Table 18 reports the category-level results on MME-RealWorld-RS. Self-directed gaze achieves the highest accuracy in all three categories. Compared with the strongest external observer in each category, GazeEarth improves Color by 7.57 percentage points, Count by 2.45 points, and Position by 4.14 points. Overall accuracy reaches 52.43%, exceeding SigLIP2 by 4.94 points.

On LRS-GRO (Table 19), GazeEarth achieves the highest Global and Object accuracies, at 77.96% and 58.04%, respectively. Its Object score exceeds RemoteCLIP by 1.80 points and SigLIP2 by 2.21 points. RemoteCLIP achieves the highest Region accuracy at 50.78%, followed by GazeEarth at 50.39%. Across the three categories, self-directed gaze yields the highest overall score of 61.00%, compared with 60.10% for RemoteCLIP and 59.54% for SigLIP2. On XLRS-Bench (Table 20), GazeEarth reaches 53.51%, exceeding SigLIP2 and RemoteCLIP by 7.41 and 7.57 percentage points, respectively.

Table 18: Observer substitution on MME-RealWorld-RS with Qwen3-VL-8B as the fixed answerer. Best and second-best marks are assigned per column over the whole table; tied values receive the same mark.

Observer	Color	Count	Position	Avg.	Δ vs. Random
Random	52.43	16.80	66.98	45.64	0.00
SigLIP2	56.10	19.33	66.35	47.49	+1.85
RemoteCLIP	55.46	18.11	66.98	47.08	+1.44
GazeEarth	**63.67**	**21.78**	**71.12**	**52.43**	**+6.79**

D.3 Effect of Gaze Depth

We evaluate gaze depth through recursive reselection and refocusing of the current view. Among these configurations, a single gaze achieves the highest benchmark-averaged accuracy with the lowest runtime. In Table 7, GazeEarth reaches 55.65% benchmark-averaged accuracy, compared with 52.91% for Two-Gaze and 52.72% for Three-Gaze. Meanwhile, average inference time increases from 5.07 seconds to 5.95 and 7.02 seconds, respectively (Table 22). The single-gaze configuration also achieves the highest overall accuracy on each of the three benchmarks among these depth variants.

Table 19: Observer substitution on LRS-GRO with Qwen3-VL-8B as the fixed answerer.

Observer	Global	Region	Object	Avg.	Δ vs. Random
Random	77.84	48.78	54.95	58.97	–
SigLIP2	77.58	49.52	55.83	59.54	+0.58
RemoteCLIP	77.75	**50.78**	56.24	60.10	+1.13
GazeEarth	**77.96**	50.39	**58.04**	**61.00**	**+2.03**

Table 20: Observer substitution with Qwen3-VL-8B as the fixed answerer on XLRS-Bench. Ties are broken using the higher overall average.

Observer	Cnt	SC	OSR	OP	Plan	AR	CR	SR	Avg.	Δ vs. Random
Random	35.00	61.33	30.60	44.10	39.00	**78.00**	65.50	43.33	45.42	0.00
SigLIP2	36.25	62.67	31.40	45.12	39.00	73.00	65.50	41.67	46.10	+0.68
RemoteCLIP	37.50	61.67	29.80	45.12	38.00	77.00	65.50	43.33	45.94	+0.52
GazeEarth	**44.38**	**66.67**	**37.80**	**53.07**	**49.00**	76.00	**74.50**	**55.00**	**53.51**	**+8.09**

In the repeated-selection variants, each subsequent gaze selects regions from an indexed overview of the already warped scene and applies another focus transformation to that view. The transformations therefore compose: each new request reallocates the current canvas, and a shift in focus can compress regions enlarged by an earlier round. The answerer receives the final transformed view rather than an accumulated collection of observations. Repeated selection thus revises the spatial allocation within the same canvas. The single-gaze configuration instead uses one spatial request to construct the full-scene observation directly from the original image, achieving the strongest accuracy–runtime trade-off among the evaluated gaze depths.

D.4 Random Focus versus Overview

Random Focus applies the same full-scene focus operator to randomly selected regions. Its benchmark-averaged accuracy is 50.01%, compared with 51.38% for Overview (Tables 6 and 7). On LRS-GRO, accuracy decreases from 59.99% to 58.97%; on MME-RealWorld-RS, it decreases from 49.01% to 45.64%. XLRS-Bench increases from 45.13% to 45.42%. The LRS-GRO breakdown (Table 16) places the largest reductions in Region and Object questions, which fall from 50.22% to 48.78% and from 56.22% to 54.95%, respectively. Under a fixed canvas budget, enlarging a region compresses the periphery, making the location of the allocated pixels consequential for answering.

With the focus operator and answerer held fixed, SigLIP2 and RemoteCLIP each achieve a three-benchmark mean accuracy of 51.04%, compared with 50.01% for random selection. Self-directed gaze reaches 55.65% and is the only evaluated observer configuration to exceed Overview's mean accuracy of 51.38%, by 4.27 percentage points.

E Statistical Analysis

To rigorously evaluate the performance improvements, we conducted paired image-cluster label-swap permutation tests across the three benchmarks to calculate p-values, while 95% confidence intervals (CIs) were estimated via paired bootstrap. To control the family-wise error rate across the nine primary comparisons, all reported p-values are adjusted using the Holm-Bonferroni method (p_{Holm}). As shown in Table 21, GazeEarth yields highly significant accuracy gains over the Direct baseline across all datasets ($p_{\text{Holm}} < 0.001$). When compared to the full-scene Overview, GazeEarth demonstrates strongly significant improvements on contextually demanding datasets, achieving a gain of +3.42 percentage points on MME-RealWorld-RS ($p_{\text{Holm}} < 0.001$) and a substantial +8.38 points on XLRS-Bench ($p_{\text{Holm}} < 0.001$). On LRS-GRO, where the uniform Overview strategy already reaches a high performance threshold (59.99%), the additional gain from GazeEarth (+1.01 points) is comparatively marginal and not statistically significant after adjustment ($p_{\text{Holm}} = 0.078$). These results indicate that GazeEarth reliably extracts critical local evidence to overcome resolution bottlenecks, while functioning as a stable, non-degrading observation mechanism when global views are already sufficient.

Table 21: **Statistical significance of paired comparisons.** Δ indicates the accuracy difference (percentage points) between the first and second method. The 95% confidence intervals (CIs) are derived from paired image-cluster bootstrap. p-values are computed via paired image-cluster label-swap tests and adjusted across the nine comparisons using the Holm-Bonferroni method (p_{Holm}).

Benchmark	Comparison	Δ	95% CI	p_{Holm}
XLRS-Bench	Overview − Direct	$+2.21$	$[+0.85, +3.54]$	0.004
	GazeEarth − Direct	$+10.59$	$[+9.05, +12.13]$	< 0.001
	GazeEarth − Overview	$+8.38$	$[+6.85, +9.91]$	< 0.001
LRS-GRO	Overview − Direct	$+3.09$	$[+2.15, +4.03]$	< 0.001
	GazeEarth − Direct	$+4.10$	$[+3.08, +5.12]$	< 0.001
	GazeEarth − Overview	$+1.01$	$[-0.12, +2.14]$	0.078
MME-RealWorld-RS	Overview − Direct	$+6.90$	$[+5.11, +8.69]$	< 0.001
	GazeEarth − Direct	$+10.32$	$[+8.45, +12.19]$	< 0.001
	GazeEarth − Overview	$+3.42$	$[+1.85, +4.99]$	< 0.001

F Efficiency Analysis

Table 22 reports end-to-end runtime across the three benchmarks. Tables 23– 25 decompose this runtime into selection, visual construction, and answering, using the protocol in Appendix A.4.

Table 22: Qwen3-VL-8B inference efficiency. Runtime is measured in seconds; lower is better. Speedup is relative to Direct; higher is better.

Variant	XLRS-Bench (s) ↓	LRS-GRO (s) ↓	MME-RealWorld-RS (s) ↓	Avg. (s) ↓	Speedup ↑
Visual Evidence Composition					
Direct	8.051	6.417	7.563	7.344	–
Overview	4.096	**3.967**	**3.998**	**4.020**	**1.827**×
Random Focus	4.687	5.124	5.019	4.943	1.486×
Gaze + Crop	**3.945**	4.689	4.661	4.432	1.657×
Think with Image	4.690	5.323	5.088	5.034	1.459×
Linked Global + Crop	4.702	5.307	5.365	5.125	1.433×
AttWarp	4.306	4.786	4.209	4.434	1.656×
Gaze Depth and Iterative Baselines					
Two-Gaze	5.753	6.170	5.936	5.953	1.234×
Three-Gaze	6.835	7.298	6.931	7.021	1.046×
GazeEarth	4.789	5.135	5.272	5.065	1.450×

Table 23: Qwen3-VL-8B inference efficiency on MME-RealWorld-RS. Lower is better for the number of MLLM calls and runtime components, while higher is better for speedup. Ties are broken using the lower total runtime.

Variant	MLLM calls ↓	Selection (s) ↓	Visual construction (s) ↓	Answer (s) ↓	Total (s) ↓	Speedup vs. Direct ↑
Visual Evidence Composition						
Direct	1	–	–	7.563	7.563	–
Overview	1	–	**0.147**	3.851	**3.998**	**1.892**×
Random Focus	2	0.889	0.325	3.805	5.019	1.507×
Gaze + Crop	2	0.905	0.176	**3.580**	4.661	1.623×
Think with Image	2	**0.884**	0.252	3.952	5.088	1.486×
Linked Global + Crop	2	0.986	0.305	4.075	5.365	1.410×
AttWarp	2	–	–	–	4.209	1.797×
Gaze Depth and Iterative Baselines						
Two-Gaze	3	1.763	0.370	3.803	5.936	1.274×
Three-Gaze	4	2.683	0.445	3.803	6.931	1.091×
GazeEarth	≤ 2	1.011	0.355	3.907	5.272	1.435×

F.1 Model calls and observation construction.

GazeEarth uses one question-guided selection call followed by one answering call; a directly resolved spatial request bypasses selection. The focus operator constructs the observation between these calls through deterministic resampling. Selection takes 1.011 seconds on MME-RealWorld-RS and 1.125 seconds on LRS-GRO. Visual construction, including focused-view rendering, takes 0.355 and 0.210 seconds, respectively, accounting for 6.7% and 4.1% of total runtime. The selection call therefore contributes most of the additional cost over Overview.

Table 24: Qwen3-VL-8B inference efficiency on LRS-GRO. Lower is better for the number of MLLM calls and runtime components, while higher is better for speedup. Ties are broken using the lower total runtime.

Variant	MLLM calls ↓	Selection (s) ↓	Visual construction (s) ↓	Answer (s) ↓	Total (s) ↓	Speedup vs. Direct ↑
Visual Evidence Composition						
Direct	1	–	–	6.417	6.417	–
Overview	1	–	0.169	3.798	**3.967**	**1.618**×
Random Focus	2	1.118	0.212	3.794	5.124	1.252×
Gaze + Crop	2	1.141	**0.060**	**3.488**	4.689	1.369×
Think with Image	2	1.155	0.192	3.976	5.323	1.206×
Linked Global + Crop	2	**0.970**	0.364	3.973	5.307	1.209×
AttWarp	2	–	–	–	4.786	1.341×
Gaze Depth and Iterative Baselines						
Two-Gaze	3	2.117	0.249	3.804	6.170	1.040×
Three-Gaze	4	3.195	0.288	3.815	7.298	0.879×
GazeEarth	≤ 2	1.125	0.210	3.800	5.135	1.250×

Table 25: Qwen3-VL-8B inference efficiency on XLRS-Bench. Lower is better for the number of MLLM calls and runtime components, while higher is better for speedup. Ties are broken using the lower total runtime.

Variant	MLLM calls ↓	Selection (s) ↓	Visual construction (s) ↓	Answer (s) ↓	Total (s) ↓	Speedup vs. Direct ↑
Visual Evidence Composition						
Direct	1	–	–	8.051	8.051	–
Overview	1	–	0.331	3.765	4.096	1.966×
Random Focus	2	**0.459**	0.481	3.747	4.687	1.718×
Gaze + Crop	2	0.465	**0.193**	**3.287**	**3.945**	**2.041**×
Unlinked Global + Crop	2	0.474	0.422	3.794	4.690	1.717×
Linked Global + Crop	2	0.483	0.421	3.798	4.702	1.712×
AttWarp	2	–	–	–	4.306	1.870×
Gaze Depth and Iterative Baselines						
Two-Gaze	3	1.427	0.570	3.756	5.753	1.399×
Three-Gaze	4	2.430	0.654	3.751	6.835	1.178×
GazeEarth	≤ 2	0.474	0.532	3.783	4.789	1.681×

F.2 Answering cost.

Answer generation accounts for approximately three quarters of GazeEarth's total runtime. It takes 3.907 seconds on MME-RealWorld-RS and 3.800 seconds on LRS-GRO, close to Overview's 3.851 and 3.798 seconds. Compared with Direct, the answering stage is shorter by 3.656 and 2.617 seconds, respectively. These reductions exceed the combined selection and visual-construction costs of 1.366 and 1.335 seconds, producing the overall runtime advantage reported in the main text.

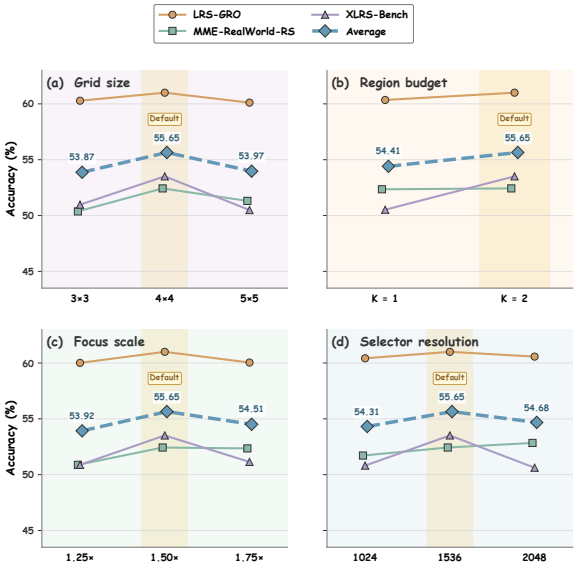

Figure 8: Hyperparameter sensitivity of GazeEarth with Qwen3-VL-8B: (a) selection-grid size, (b) maximum number of selected cells, (c) focus enlargement factor, and (d) selector-overview long-side cap in pixels. Solid curves show individual benchmark scores; the dashed curve shows their unweighted mean. Shaded bands mark the default settings.

G HYPERPARAMETER SENSITIVITY

Figure 8 examines four settings that control spatial selection and pixel allocation. The default configuration uses a 4×4 grid, a maximum of two selected cells, a focus enlargement factor of $\alpha = 1.5$, and a selector-overview long-side cap of 1536 pixels. Tables 26 – 28 provide category-level results on MME-RealWorld-RS, LRS-GRO, and XLRS-Bench.

G.1 SELECTION-GRID SIZE.

The 4×4 grid achieves a benchmark-averaged accuracy of 55.65%, compared with 53.87% for 3×3 and 53.97% for 5×5. It also gives the highest overall score on each benchmark among the three grid settings. Increasing selection granularity beyond 4×4 therefore does not improve answering performance.

G.2 REGION BUDGET.

Increasing the maximum number of selected cells from one to two raises mean accuracy from 54.41% to 55.65%. On LRS-GRO, Object accuracy increases from 56.69% to 58.04%, while Global accuracy increases from 77.84% to 77.96%. A two-cell budget allows a spatial request to cover separated evidence within a single selection round.

G.3 FOCUS SCALE.

The enlargement factor controls how much output area is assigned to the selected neighbourhood. Mean accuracy reaches 55.65% at $\alpha = 1.5$, compared with 53.92% at 1.25 and 54.51% at 1.75. The category-level results reveal different responses to stronger magnification. On MME-RealWorld-RS, increasing α from 1.5 to 1.75 improves Color and Count accuracy but reduces Position accuracy from 71.12% to 68.10%. The intermediate scale yields the highest cross-benchmark mean.

G.4 SELECTOR RESOLUTION.

A 1536-pixel overview achieves a mean accuracy of 55.65%, compared with 54.31% at 1024 pixels and 54.68% at 2048 pixels. The higher-resolution overview improves MME-RealWorld-RS accuracy from 52.43% to 52.84%, whereas LRS-GRO decreases from 61.00% to 60.57%. The default resolution thus provides the highest mean score across the three benchmarks.

Table 26: Category-level parameter sensitivity on MME-RealWorld-RS. Bold setting names mark the default configuration. Best and second-best marks are assigned within each sweep and column; tied values receive the same mark.

Setting	Color	Count	Position	Avg.
Grid-size Sensitivity				
Grid 3×3	61.75	20.88	67.78	50.37
Grid 4×4	**63.67**	**21.78**	**71.12**	**52.43**
Grid 5×5	62.15	**21.78**	69.29	51.31
Region-budget Sensitivity				
$K = 1$	**63.75**	**22.02**	70.56	52.35
$K = 2$	63.67	21.78	**71.12**	**52.43**
Focus-scale Sensitivity				
Focus 1.25	59.36	20.55	**72.00**	50.88
Focus 1.50	63.67	21.78	71.12	**52.43**
Focus 1.75	**64.62**	**23.65**	68.10	52.35
Selector-resolution Sensitivity				
Selector 1024	62.71	21.94	69.77	51.71
Selector 1536	63.67	21.78	71.12	52.43
Selector 2048	**64.06**	**22.43**	**71.28**	**52.84**

Table 27: Category-level parameter sensitivity on LRS-GRO.

Setting	Global	Region	Object	Avg.
Grid-size Sensitivity				
Grid 3×3	77.23	**50.44**	56.99	60.28
Grid 4×4	**77.96**	50.39	**58.04**	**61.00**
Grid 5×5	77.45	49.78	56.85	60.11
Region-budget Sensitivity				
$K = 1$	77.84	**50.74**	56.69	60.35
$K = 2$	**77.96**	50.39	**58.04**	**61.00**
Focus-scale Sensitivity				
Focus 1.25	77.75	50.17	56.34	60.01
Focus 1.50	**77.96**	**50.39**	**58.04**	**61.00**
Focus 1.75	77.32	50.22	56.58	60.04
Selector-resolution Sensitivity				
Selector 1024	77.66	50.83	56.87	60.42
Selector 1536	**77.96**	50.39	**58.04**	**61.00**
Selector 2048	77.79	**51.18**	56.95	60.57

Table 28: Category-level parameter sensitivity on XLRS-Bench.

Setting	Cnt	SC	OSR	OP	Plan	AR	CR	SR	Avg.
Grid-size Sensitivity									
Grid 3×3	43.13	**67.67**	34.00	50.42	45.00	77.00	70.50	46.67	50.97
Grid 4×4	**44.38**	66.67	**37.80**	**53.07**	**49.00**	76.00	**74.50**	**55.00**	**53.51**
Grid 5×5	41.25	65.33	34.20	49.88	47.00	**78.00**	70.50	46.67	50.49
Region-budget Sensitivity									
$K = 1$	40.63	**67.00**	32.80	50.06	45.00	**79.00**	71.00	48.33	50.52
$K = 2$	**44.38**	66.67	**37.80**	**53.07**	**49.00**	76.00	**74.50**	**55.00**	**53.51**
Focus-scale Sensitivity									
Focus 1.25	41.88	67.67	33.60	50.06	47.00	**78.00**	72.50	46.67	50.88
Focus 1.50	**44.38**	66.67	**37.80**	**53.07**	**49.00**	76.00	**74.50**	**55.00**	**53.51**
Focus 1.75	41.25	**68.33**	34.20	50.84	47.00	77.00	69.00	45.00	51.14
Selector-resolution Sensitivity									
Selector 1024	40.00	**67.33**	33.80	50.36	45.00	76.00	73.00	45.00	50.81
Selector 1536	**44.38**	66.67	**37.80**	**53.07**	**49.00**	76.00	**74.50**	**55.00**	**53.51**
Selector 2048	41.25	66.67	32.00	50.30	47.00	**80.00**	72.00	45.00	50.62

H Qualitative Examples

This section illustrates how spatial requests guide observation construction and subsequent answering. The examples cover spatial responses, region selection, and the resulting visual observations across methods.

H.1 Spatial Response Visualization

Figures 9–14 compare spatial responses, recorded region selections, and final answers for the same image–question pairs. The examples show how different observers distribute their responses across the scene and which regions they ultimately select. Displaying the answers alongside these selections connects where each observer looks with the resulting prediction. In each visualization, *Rec.* identifies the recorded region selection, while the cell values and rank labels describe the displayed spatial responses. The recorded selection and the response distribution are shown separately to distinguish the region used for observation construction from the spatial pattern visualized over the scene.

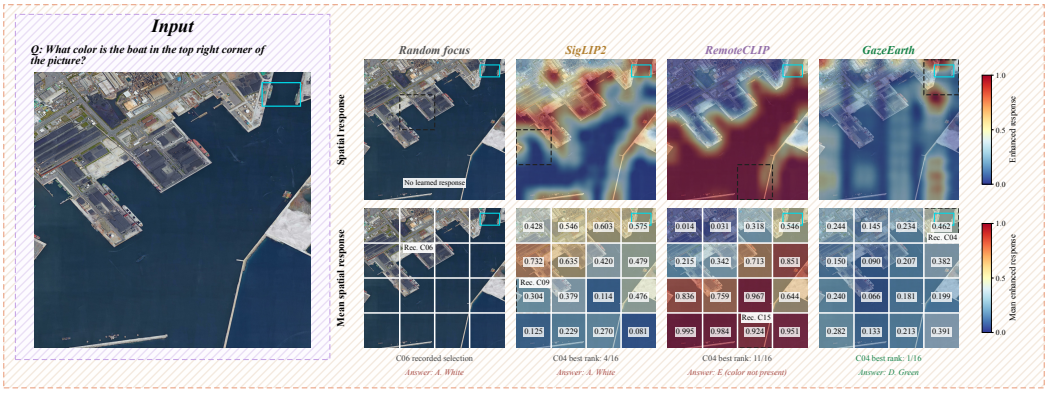

Figure 9: Spatial responses and region selections under observer substitution (Part 1). Each example compares Random Focus, SigLIP2, RemoteCLIP, and GazeEarth for the same image and question. Spatial-response maps and cell-level summaries are shown alongside the recorded selections and final answers. Random Focus selects a cell without a learned response.

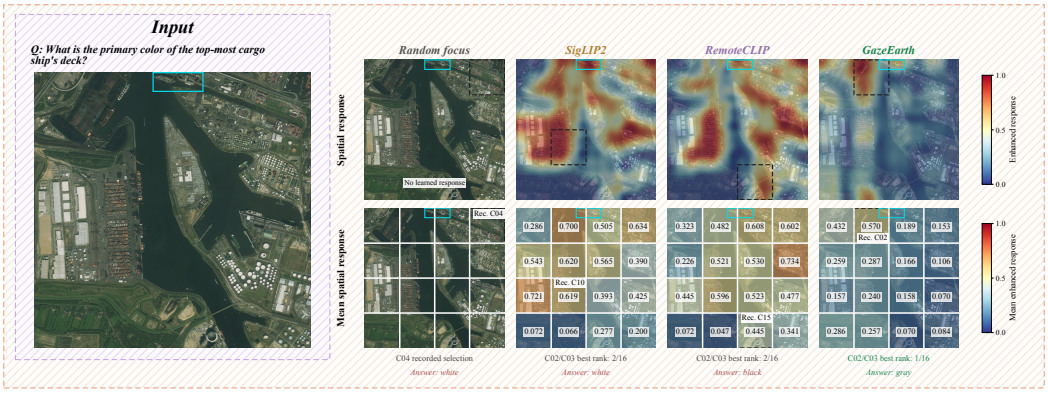

Figure 10: Spatial responses and region selections under observer substitution (Part 2). Each example compares Random Focus, SigLIP2, RemoteCLIP, and GazeEarth for the same image and question. Spatial-response maps and cell-level summaries are shown alongside the recorded selections and final answers. Random Focus selects a cell without a learned response.

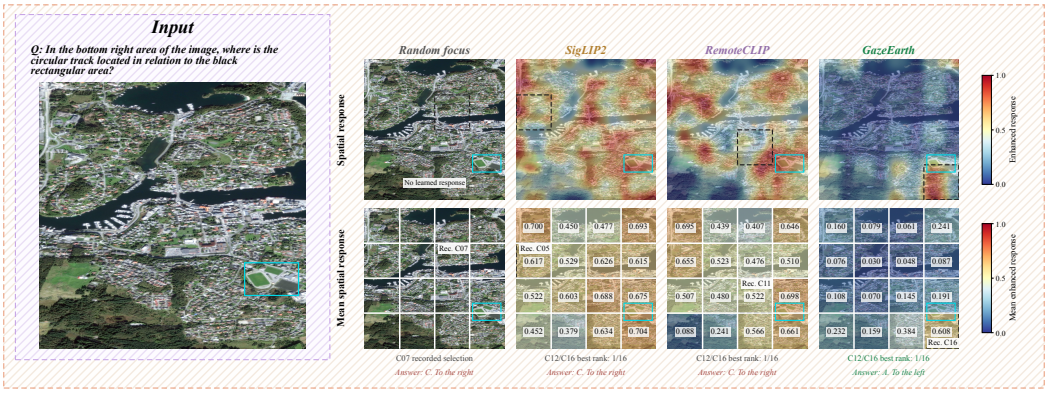

Figure 11: Spatial responses and region selections under observer substitution (Part 3). Each example compares Random Focus, SigLIP2, RemoteCLIP, and GazeEarth for the same image and question. Spatial-response maps and cell-level summaries are shown alongside the recorded selections and final answers. Random Focus selects a cell without a learned response.

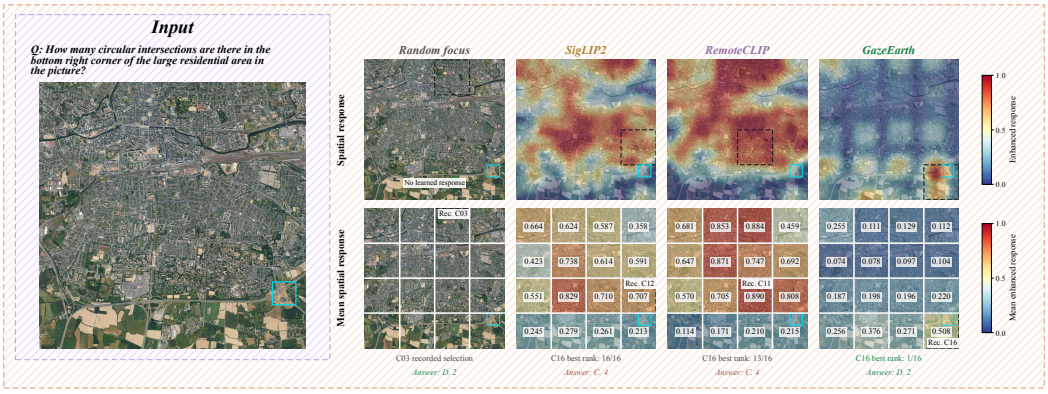

Figure 12: Spatial responses and region selections under observer substitution (Part 4). Each example compares Random Focus, SigLIP2, RemoteCLIP, and GazeEarth for the same image and question. Spatial-response maps and cell-level summaries are shown alongside the recorded selections and final answers. Random Focus selects a cell without a learned response.

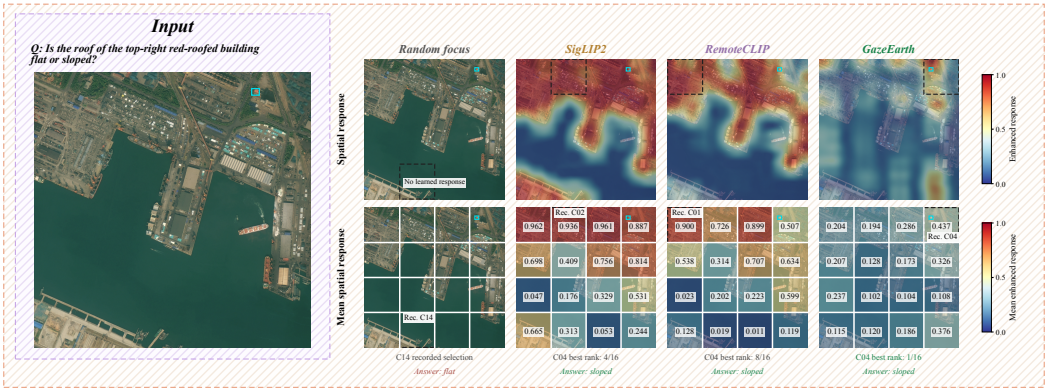

Figure 13: Spatial responses and region selections under observer substitution (Part 5). Each example compares Random Focus, SigLIP2, RemoteCLIP, and GazeEarth for the same image and question. Spatial-response maps and cell-level summaries are shown alongside the recorded selections and final answers. Random Focus selects a cell without a learned response.

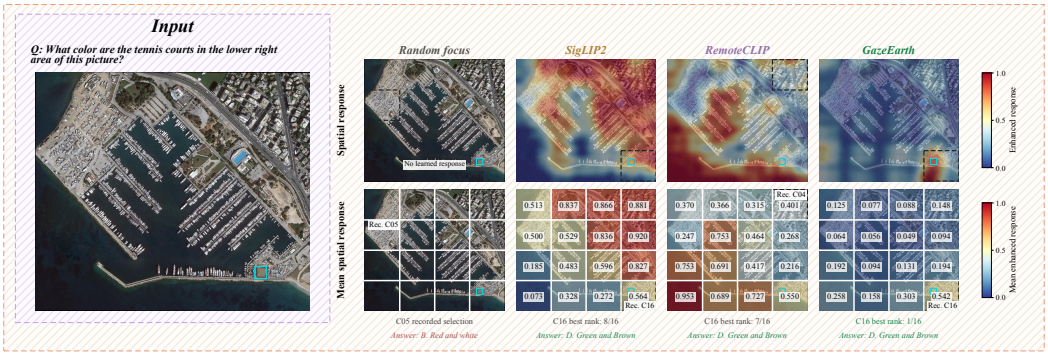

Figure 14: Spatial responses and region selections under observer substitution (Part 6). Each example compares Random Focus, SigLIP2, RemoteCLIP, and GazeEarth for the same image and question. Spatial-response maps and cell-level summaries are shown alongside the recorded selections and final answers. Random Focus selects a cell without a learned response.

H.2 Observation and Answering Examples

Figures 15–20 compare the visual observations and answers across methods on questions involving object attributes, counting, and spatial relations. GazeEarth enlarges the selected neighbourhood while retaining surrounding landmarks within the same full-scene view. These examples illustrate how local evidence and its spatial context are presented together for answering. The accompanying explanations summarize the observation and answering workflows rather than a single continuous model response. For GazeEarth, region selection produces a spatial request, and answering is performed in a separate interaction using the rendered view and the original question. Spatial preservation refers to the ordering and connectivity of scene regions under the warp; their pixel coordinates and local scales change. The examples include both shared successes across methods and differing answers to the same question, illustrating how the observation strategies behave across individual scenes.

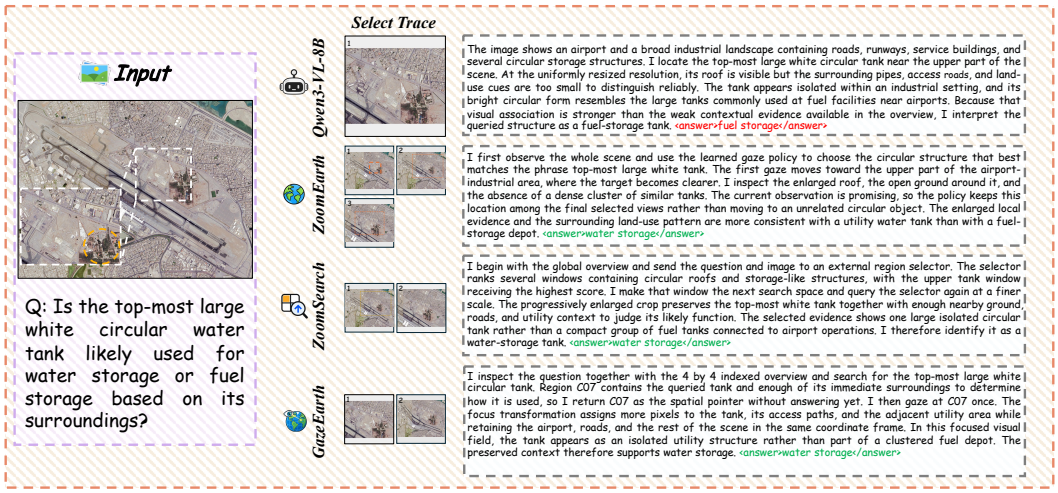

Figure 15: Qualitative comparison of observation and answering across methods (Part 1). Each example presents the input image and question, the visual observations used by the compared methods, and their accompanying explanations and final answers. GazeEarth enlarges the question-selected neighbourhood within a continuous full-scene view, retaining the surrounding spatial context for answering.

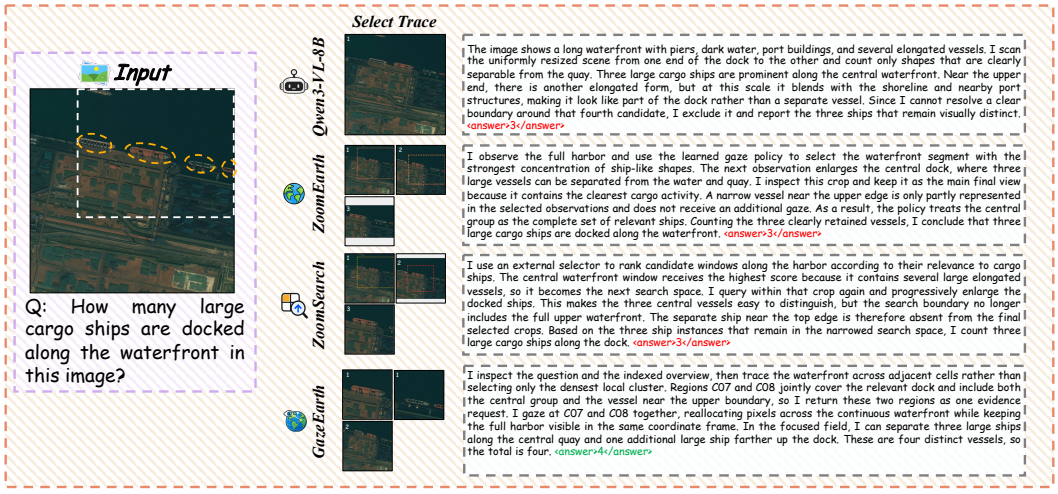

Figure 16: Qualitative comparison of observation and answering across methods (Part 2). Each example presents the input image and question, the visual observations used by the compared methods, and their accompanying explanations and final answers. GazeEarth enlarges the question-selected neighbourhood within a continuous full-scene view, retaining the surrounding spatial context for answering.

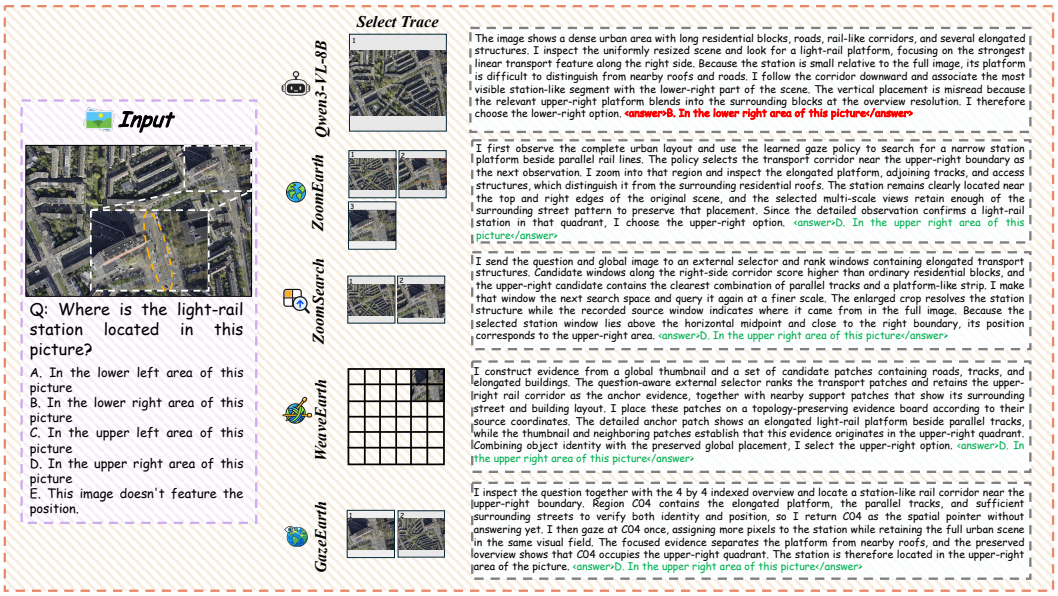

Figure 17: Qualitative comparison of observation and answering across methods (Part 3). Each example presents the input image and question, the visual observations used by the compared methods, and their accompanying explanations and final answers. GazeEarth enlarges the question-selected neighbourhood within a continuous full-scene view, retaining the surrounding spatial context for answering.

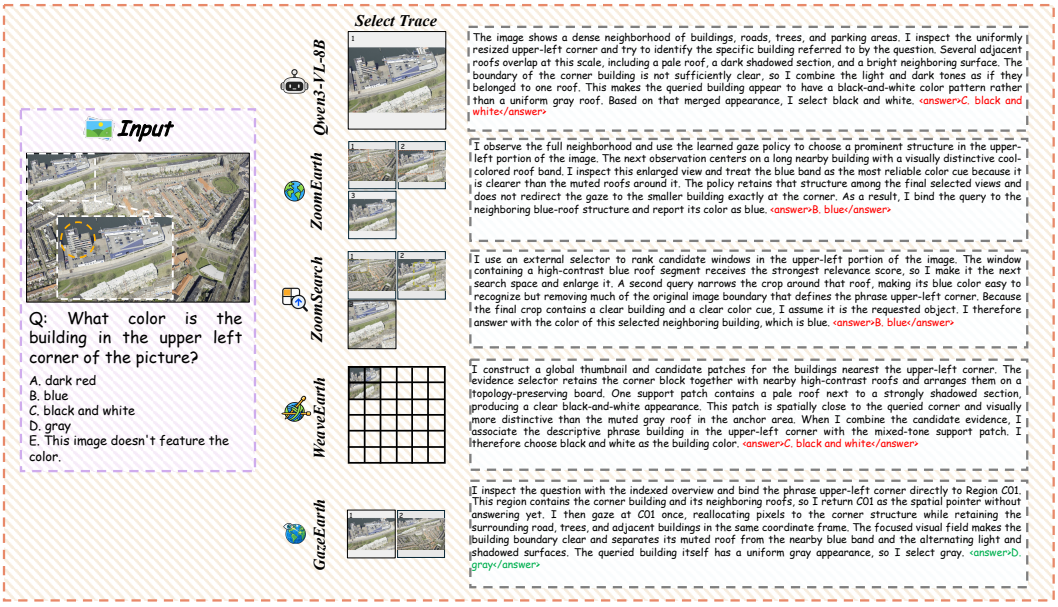

Figure 18: Qualitative comparison of observation and answering across methods (Part 4). Each example presents the input image and question, the visual observations used by the compared methods, and their accompanying explanations and final answers. GazeEarth enlarges the question-selected neighbourhood within a continuous full-scene view, retaining the surrounding spatial context for answering.

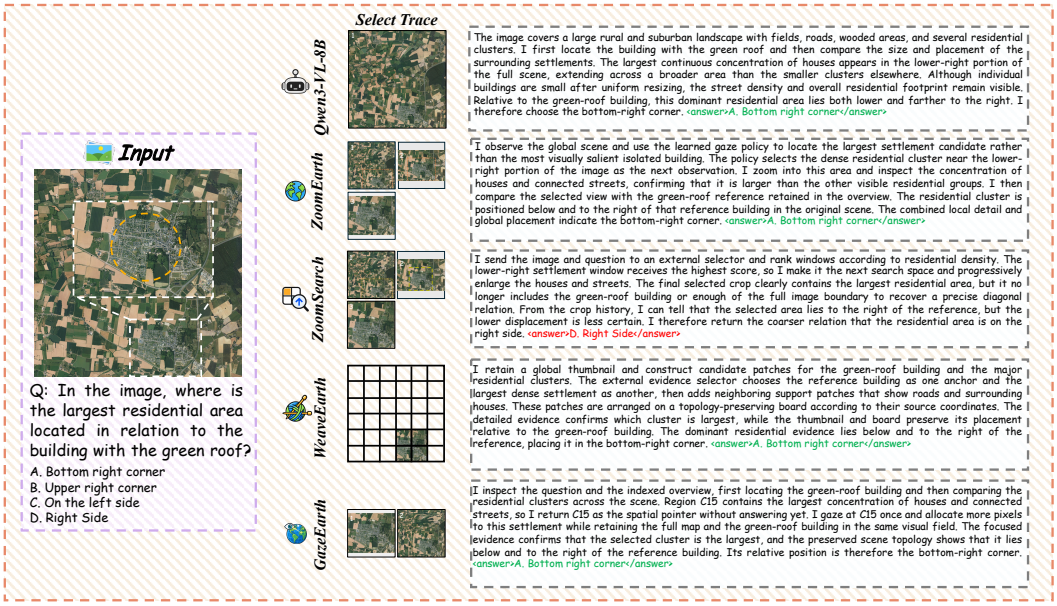

Figure 19: Qualitative comparison of observation and answering across methods (Part 5). Each example presents the input image and question, the visual observations used by the compared methods, and their accompanying explanations and final answers. GazeEarth enlarges the question-selected neighbourhood within a continuous full-scene view, retaining the surrounding spatial context for answering.

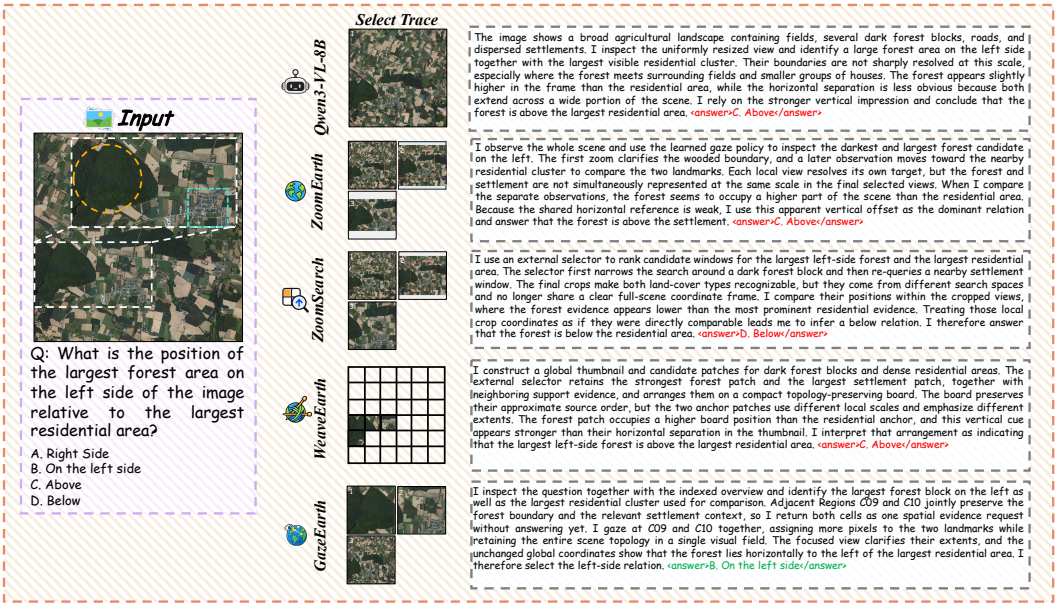

Figure 20: Qualitative comparison of observation and answering across methods (Part 6). Each example presents the input image and question, the visual observations used by the compared methods, and their accompanying explanations and final answers. GazeEarth enlarges the question-selected neighbourhood within a continuous full-scene view, retaining the surrounding spatial context for answering.

I BROADER IMPACT

UHR remote sensing supports disaster response, environmental and agricultural monitoring, infrastructure inspection, and urban planning, where analysts must relate small objects to their surrounding scene. GazeEarth improves such analysis without training or fine-tuning: it reuses frozen, publicly available MLLMs and adds only a deterministic resampling step. This lowers the barrier for groups that cannot collect task-specific annotations or train large models, such as public agencies, non-governmental organizations, and researchers in low-resource settings. Because the answer-stage canvas is bounded, GazeEarth also answers $1.45\times$ faster than native-resolution Direct inference in our Qwen3-VL-8B setting (Appendix F), reducing the compute and energy cost per query.

Unlike methods that rely on internal attention or learned zooming policies, GazeEarth exposes its spatial decision as an explicit list of grid cells, and the rendered view that the answerer receives can be inspected directly. Users can therefore check where the model looked before trusting an answer, which supports human oversight and error auditing in applied settings.

Beyond remote sensing, our results point to observation construction as a design axis for multimodal models. Current MLLMs mostly receive a fixed, uniformly resized view, and progress on high-resolution inputs is usually sought through larger encoders, longer visual contexts, or additional training. Our findings suggest a complementary route: a frozen model can already express where it needs more detail, and how that request is turned into the next observation affects what the model can answer, sometimes as much as the choice of region itself.