TRACEEARTH: FROM SEMANTIC MATCHING TO SPATIAL TRACING FOR TRAINING-FREE UHR REMOTE SENSING VQA

Anonymous authors
Paper under double-blind review

ABSTRACT

Ultra-high-resolution (UHR) remote sensing visual question answering (VQA) requires capturing global context and local details within a limited visual token budget, making efficient acquisition of question-relevant local evidence critical. Existing methods improve UHR perception through visual token compression or selection, question-guided region retrieval, or iterative zooming and search. However, these methods mainly determine what visual content to retain or inspect based on semantic relevance or visual importance, while explicit spatial cues in the question are rarely converted into direct constraints on the search space. This can lead to redundant visual processing, localization errors, and missed critical evidence. We argue that remote sensing questions should not only identify "which regions are relevant," but directly indicate "where the target is." Converting spatial cues into executable visual search priors shifts broad semantic search toward efficient, precise target localization. We therefore propose TraceEarth, a training-free framework for spatially guided visual evidence acquisition in UHR remote sensing VQA. TraceEarth decouples spatial cues from full-question semantics, shifting visual evidence acquisition from semantic matching to spatial tracing. Building on this framework, we further propose a spatial cue guided evidence tracing method that combines image-referenced localization, object-referenced tracing, relation-preserving evidence construction, and candidate verification using frozen models and existing visual tools, progressively acquiring question-relevant evidence from global scenes to local targets. TraceEarth achieves 60.03% and 61.10% accuracy on the UHR benchmarks MME-RealWorld-RS and XLRS-Bench Lite, respectively, and 72.61% on the remote sensing benchmark LHRS-Bench. Consistent gains across base vision-language models demonstrate its effectiveness and generalizability without additional training. Our code will be released.

1 INTRODUCTION

Ultra-high-resolution (UHR) remote sensing visual question answering (VQA) requires reasoning over geographic objects, attributes, counts, and spatial relationships in large-scale imagery. (Wang et al., 2025a; Zhang et al., 2025c; Wang et al., 2026a; Ma et al., 2026b) Due to extensive scene coverage and large variations in object scale (Wang et al., 2026a), question-relevant evidence often occupies only a small region, (Liu et al., 2025; Ma et al., 2026b) making it difficult to preserve both global context and fine-grained details within a limited visual budget. Efficiently localizing relevant regions is therefore a key challenge in UHR remote sensing VQA (Wang et al., 2026a;b). To reduce visual redundancy in UHR images, existing methods mainly employ visual token compression or selection (Luo et al., 2025; Dang et al., 2026; Li et al., 2026c; Ma et al., 2026a), semantic region retrieval (Wang et al.), or progressive visual search (Wang et al., 2026a; Liu et al., 2026b; Zhang et al., 2025b; Wang et al., 2026b). Token-based approaches prune, merge, or selectively retain informative visual tokens to reduce the effective input size, while retrieval-based approaches partition high-resolution images into local regions and identify question-relevant evidence according to cross-modal semantic relevance. Search-based approaches instead progressively acquire finer visual evidence through cropping, zooming, or iterative interaction. Together, these methods demonstrate

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

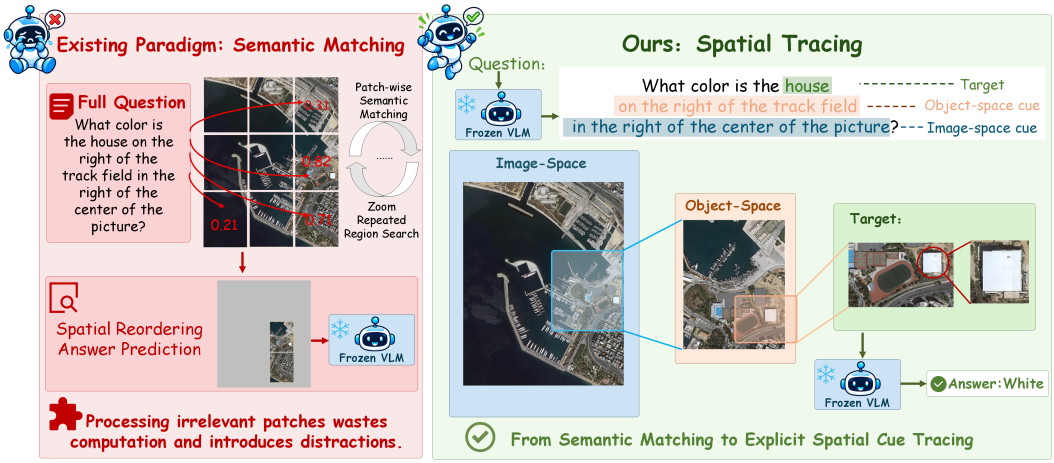


Figure 1: Motivation behind TraceEarth. Spatial cues that support target localization are prevalent in UHR remote sensing VQA. Rather than treating the full question as a unified semantic query for region matching, TraceEarth decouples spatial cues from question semantics and constructs spatial tracing chains based on dependencies across image space and object space, shifting from semantic matching to spatial tracing.

that selectively allocating visual computation to task-relevant evidence is critical for efficient UHR remote sensing understanding.

Despite their success, existing methods guide visual search by semantic relevance between the question and regions, requiring comparisons across multiple candidates. Yet, as shown in Figure 2, over 90% of questions in UHR remote sensing VQA contain spatial cues describing target locations or spatial relationships between objects. Since repeated instances of the same category often appear across large scenes, category or attribute semantics alone may not uniquely identify the target region. Spatial cues therefore provide constraints for narrowing the search space and resolving target ambiguity. These spatial cues guide visual search in distinct ways. For example, “in the upper right area of the image ” specifies a target region in the image, whereas “ the building to the right of the track field” defines the target relative to a reference object. Treating both types of cues together with the full question semantics in a unified retrieval process weakens their explicit spatial guidance and requires the vision language model to jointly perform region localization, reference recognition, and spatial relation reasoning. We therefore argue that spatial information should directly guide “where the target is,” rather than merely indicate “which regions are relevant.” By decoupling spatial cues from question semantics and converting them into executable constraints, visual search can progressively trace the target and acquire local evidence. In this work, we propose TraceEarth, a training-free framework for spatially guided visual evidence acquisition in UHR remote sensing VQA. TraceEarth decouples spatial cues from the full question semantics and organizes them into tracing chains that follow spatial dependencies from global scenes to local targets, transforming spatial cues into executable constraints for evidence acquisition. Within this framework, we develop a spatial cue guided evidence tracing method. It combines image-referenced localization, object-referenced tracing, relation-preserving evidence construction, and candidate verification using frozen models and existing visual tools. Known spatial constraints directly guide evidence tracing, while candidate locations are treated as hypotheses to be verified. In this way, TraceEarth shifts visual evidence acquisition from semantic matching to spatial tracing.

Extensive experiments across multiple remote sensing VQA benchmarks show that TraceEarth improves UHR remote sensing VQA performance without additional training or modification of the base vision language model. It can also be integrated as a plug and play module into existing visual search methods, improving accuracy while reducing search and inference costs.

- We reveal the prevalence of spatial cues in UHR remote sensing VQA and their role in visual evidence localization. Based on this observation, we propose TraceEarth, which

decouples spatial cues from the full question semantics and shifts local visual evidence acquisition from semantic matching to spatial tracing.

- We further propose a spatial cue guided evidence tracing method. It combines image-referenced localization, object-referenced tracing, relation-preserving evidence construction, and candidate verification using frozen models and existing visual tools, progressively acquiring question-relevant evidence from global scenes to local targets while reducing redundant search and missed critical evidence.
- Extensive experiments on multiple remote sensing VQA benchmarks demonstrate that our method achieves state-of-the-art performance compared with existing remote sensing VQA methods, while remaining training-free and plug-and-play and reducing visual search and inference costs.

2 Related Work

2.1 Remote Sensing Vision Language Models

Recent advances in multimodal large language models have accelerated intelligent Earth observation and driven the development of remote sensing vision language models, progressively expanding their capabilities from basic image understanding and visual question answering toward multitask and fine grained visual language interpretation (Hu et al., 2025; Muhtar et al., 2025; Zhang et al., 2024a; Zhan et al., 2025). Meanwhile, remote sensing foundation models have further advanced general representation learning across diverse Earth observation data (Guo et al., 2024; Wu et al., 2025; kelu et al., 2025). For example, GeoChat (Kuckreja et al., 2024) introduces region level interaction and visual grounding, VHM (Pang et al., 2024) enhances fine grained remote sensing understanding, EarthDial (Sagar Soni, 2025) extends multimodal interaction to diverse Earth observation data, LHRS Bot Nova (Li et al., 2024) strengthens visual language alignment and spatial understanding, and GeoCoT (Li et al., 2026a) improves reliable remote sensing reasoning. Overall, remote sensing vision language models have substantially advanced Earth observation understanding, but UHR imagery remains challenging because relevant evidence often occupies only a small region amid substantial visual redundancy and a limited visual token budget.

2.2 Efficient Visual Perception for UHR Remote Sensing

Efficient high resolution visual perception mainly follows two directions. Passive methods reduce redundant visual inputs through token compression, token selection, or region retrieval (Wang et al., 2025a; Zhang et al., 2025c; Dang et al., 2026; Luo et al., 2025). Active methods instead iteratively acquire task relevant local evidence through visual search, cropping, zooming, refocusing, or pixel space interaction (Wu & Xie, 2023; Zhang et al., 2025a; Li et al., 2025; Su et al., 2025; Li et al., 2026b; Zheng et al., 2025; Qi et al., 2024). These strategies have also been increasingly explored in UHR remote sensing (Zhang et al., 2025b; Liu et al., 2026b). For example, ZoomSearch (Zhou et al., 2025) performs hierarchical region search, ZoomEarth (Liu et al., 2026a) uses adaptive cropping and zooming, GeoEyes (Wang et al., 2026a) supports on demand visual focusing, and WeaveEarth (Ma et al., 2026b) constructs complementary global and local evidence. Despite improved efficiency, existing methods determine what or where to inspect through semantic relevance, visual feedback, learned policies, or model generated decisions. Explicit spatial cues in the question are rarely converted into executable constraints for visual evidence acquisition.

3 Observation

Experimental Setup. We analyze spatial cues in the questions of MME RealWorld RS and XLRS-Bench Lite. We define spatial cues as expressions that explicitly constrain the location of the target or related entities. Based on their spatial references, we further define two types of spatial cues: image space cues and object space cues. Image space cues use the entire image as the spatial reference and specify target locations through directional descriptions, coordinates, bounding boxes, or visual markers that can be directly mapped to image regions. Object Space cues use scene entities as spatial references and constrain targets through relative spatial relationships. The two types may coexist in the same question.

Observation. As shown in Figure 2, spatial cues are highly prevalent, appearing in 95.3% of MME RealWorld RS samples and 91.2% of XLRS-Bench Lite samples. Specifically, image space cues account for 90.0% and 87.4%, while Object Space cues account for 43.8% and 20.4%, respectively. These results reflect a common ambiguity in UHR scenes, where repeated instances of the same category make category or attribute semantics insufficient for unique localization. Image space cues narrow the target to a global region, while Object Space cues further resolve ambiguity through reference entities and spatial relations. This motivates us to decouple spatial cues from full question semantics and exploit their dependencies to guide local visual evidence acquisition.

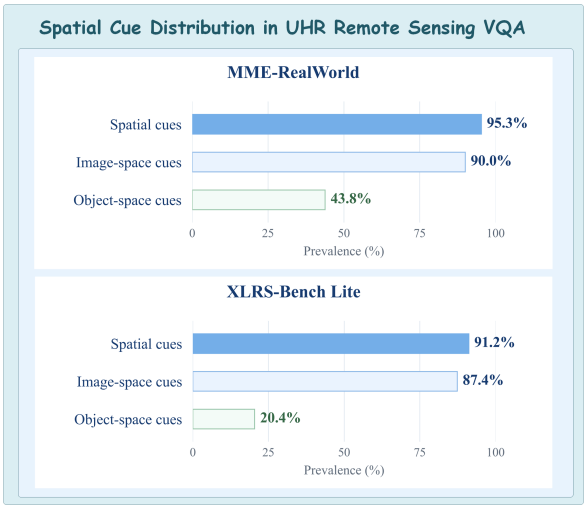


Figure 2: Spatial cue distribution in UHR remote sensing VQA. Spatial cues are widely distributed across UHR remote sensing VQA datasets, supporting our subsequent use of their spatial dependencies to guide visual evidence acquisition.

4 METHOD

4.1 OVERVIEW

As shown in Figure 3, TraceEarth is a training-free framework for spatially guided visual evidence acquisition in UHR remote sensing VQA. An Agent parses spatial cues according to their references in image space and object space, identifies their roles in the current question, and plans corresponding tracing paths. It then progressively translates image directions, entity references, spatial relations, and candidate locations into region localization, relational tracing, and local verification, tracing visual evidence from the global scene to local targets. The process consists of three stages. Spatial Intent Routing determines how spatial cues should be used based on the question and candidate answers. Image space Tracing converts image level spatial constraints into region localization and cropping, while Object Space Tracing uses target entities, reference entities, and known spatial relations to further trace relevant local content. These stages are executed as needed by the same frozen base model with task specific instructions and visual tools. The resulting local evidence is finally combined with the original question for answer reasoning.

4.2 SPATIAL INTENT ROUTING

Spatial Intent Routing determines how spatial cues should be used for subsequent evidence acquisition. Given a question Q and candidate answers $\mathcal{C}$, it analyzes their spatial semantics and outputs an execution path:

$$\mathcal{P} = \Psi(Q, \mathcal{C}), \qquad \mathcal{P} \in \{\mathcal{P}_I, \mathcal{P}_O, \mathcal{P}_G\},$$

where $\mathcal{P}_I$, $\mathcal{P}_O$, and $\mathcal{P}_G$ denote image position prediction, object spatial relation prediction, and general VQA guided by spatial conditions, respectively.

For $\mathcal{P}_I$, candidate answers denote possible target locations and are mapped to regions for verification. For $\mathcal{P}_O$, candidate spatial relations are prediction targets and cannot guide tracing; only known cues are used to localize relevant entities while preserving the layout for relation prediction. For $\mathcal{P}_G$, spatial cues are known localization conditions that directly guide tracing. Spatial Intent Routing thus determines the role of spatial information and the tracing strategy, then invokes image space tracing and object space tracing as needed. Details are provided in Appendix A.

4.3 IMAGE-SPACE TRACING

Image-Space Parsing and Region Tracing. Image-Space Tracing converts image level spatial cues into local visual regions. Given a question Q and image input I, the agent first decouples

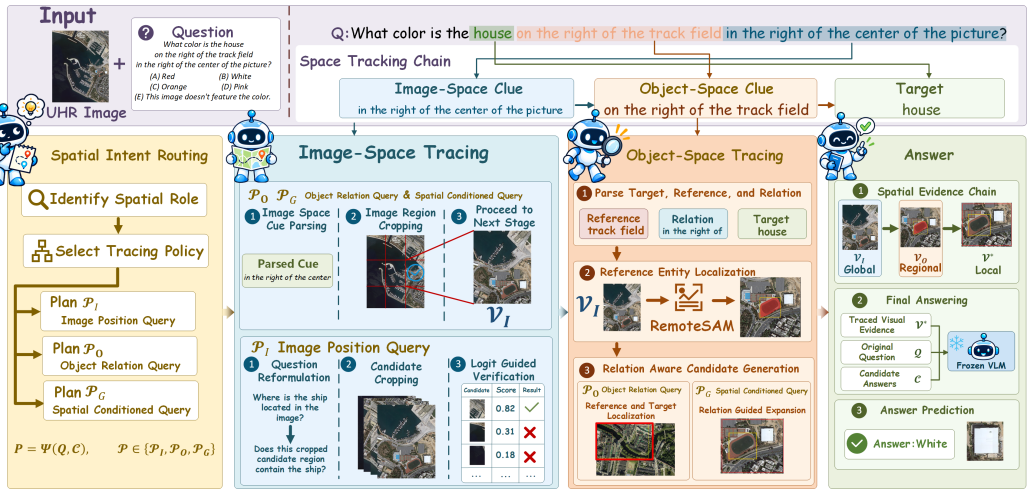


Figure 3: Overall framework of TraceEarth. The agent identifies the role of spatial cues and plans the corresponding tracing path, then performs image space and object space Tracing as needed to translate image directions, entity references, and spatial relations into region localization, candidate verification, and relation guided tracing. This progressively traces visual evidence from the global scene to local targets for final answer prediction.

image space cues as $\mathcal{S}_I = \Phi_I(Q)$, where $\Phi_I(\cdot)$ is performed by the frozen base VLM under a dedicated instruction, and $\mathcal{S}_I$ represents spatial constraints relative to the current image. These cues include natural language directions, bounding boxes, visual markers, and other representations directly mapped to image regions. We define an image space skill set $\mathcal{K}_I = \{k_I^1, k_I^2, \ldots, k_I^M\}$, including Region Localization, Bounding Box Cropping, and Visual Marker Localization. Given $\mathcal{S}_I$, the agent selects a Skill and its execution parameters:

$$(k_I^*, \theta_I) = \Psi_I(\mathcal{S}_I, \mathcal{K}_I),$$

where θ_I specifies region directions, bounding coordinates, or visual marker locations. The selected Skill is then applied to I:

$$\mathcal{V}_I = k_I^*(I; \theta_I),$$

yielding the local region for subsequent candidate verification or Object Space Tracing. Skill details are provided in Appendix B.

Candidate Conditioned Spatial Tracing. For image position questions $\mathcal{P}_I$, candidate answers denote possible target locations and are treated as tracing branches to be verified. Given a question Q, image I, and candidate c_i, the corresponding local region is obtained as $V_i = \Gamma_I(I, c_i)$, where $\Gamma_I(\cdot)$ maps the directional description in c_i to an image region and crops it from I. The original question is then reformulated as a local binary verification question $\widetilde{Q}_i = \mathcal{R}(Q, c_i)$, where $\mathcal{R}(\cdot)$ is performed by the frozen base VLM under a dedicated instruction while preserving the target category, attributes, and necessary semantics. For example, "Where is the ship located in the image?" with candidate top left is converted to "Does this cropped candidate region contain the ship?" Details are provided in Appendix B.2. We feed $(V_i, \widetilde{Q}_i)$ into the frozen base VLM and obtain the Yes and No logits z_i^+ and z_i^-. The evidence score is

$$s_i = \frac{\exp(z_i^+)}{\exp(z_i^+) + \exp(z_i^-)}.$$

The candidate with the highest score is selected. If the gap between the highest and second highest scores is below $\tau = 0.03$, the agent falls back to the original image for final prediction.

4.4 Object-Space Tracing

Object Space Parsing and Relational Tracing. Object Space Tracing uses spatial relations between entities to trace target evidence within the current visual region. Given a question Q and

Table 1: Comparison with existing methods on MME-RealWorld-RS and LHRS-Bench. The best and second-best results are highlighted in **bold** and underlined, respectively.

Method	MME-RealWorld-RS				LHRS-Bench
	Color	Count	Position	Avg.	Acc.
GPT-4o	34.18	15.17	37.07	28.81	66.19
Claude-3.7-Sonnet	32.90	19.42	28.57	27.02	–
LLaVA-v1.5-7B	23.11	16.88	26.25	22.12	47.17
Qwen2.5-VL-7B	45.94	21.05	55.23	40.90	53.98
RSUniVLM-0.5B	28.97	19.00	28.33	25.48	–
GeoChat	23.11	15.66	25.06	21.32	37.62
VHM-7B	20.32	16.80	35.24	24.18	–
ZoomEye	60.56	35.56	48.45	48.19	–
ZoomEarth	49.69	31.66	50.14	43.93	–
GeoLLaVA-8K	27.92	22.27	34.90	28.41	–
ImageRAG	60.48	32.46	63.33	52.09	–
RAP	64.53	**40.25**	57.62	54.20	–
RADAR	50.52	20.47	58.15	43.05	67.47
ZoomSearch	66.14	39.15	67.62	57.64	–
WeaveEarth	51.39	27.65	62.60	47.38	–
TraceEarth (Ours)	**69.32**	32.38	**77.72**	**60.03**	**72.61**

current visual input $\mathcal{V}_I$, the agent first decouples the Object Space information as $\mathcal{S}_O = \Phi_O(Q) = \{o_t, o_a, \mathcal{R}_k\}$, where $\Phi_O(\cdot)$ is performed by the frozen base VLM under a dedicated instruction, o_t is the target entity, o_a is the reference entity, and $\mathcal{R}_k$ is a known spatial relation in the question. If image space tracing has been performed, $\mathcal{V}_I$ is its output; otherwise, the original image is used.

The agent then localizes the reference entity as $\mathcal{B}_a = D(\mathcal{V}_I, o_a)$, where D is the entity localizer implemented with RemoteSAM. Starting from $\mathcal{B}_a$, the agent traces the region according to $\mathcal{R}_k$:

$$\mathcal{V}_O = G_O(\mathcal{V}_I, \mathcal{B}_a, \mathcal{R}_k),$$

where $G_O(\cdot)$ determines the main expansion direction while preserving necessary orthogonal context. The resulting $\mathcal{V}_O$ is the local region indicated by the relation. Implementation details are provided in Appendix B.3.

Relation Preserving Evidence Construction. For object spatial relation questions $\mathcal{P}_O$, the relation to be predicted cannot be used in advance as an Object Space tracing constraint. The agent therefore localizes relevant entities using only known entity descriptions while preserving their original spatial layout for relation reasoning.

Given the current visual region $\mathcal{V}_I$, the entity localizer $D(\cdot)$ detects relevant instances as $\mathcal{B} = \{b_1, \ldots, b_n\}$, where b_i is the bounding box of the i-th instance. We then construct the smallest enclosing region and crop the local evidence:

$$b^* = \mathrm{Envelope}(\mathcal{B}), \qquad \mathcal{V}_O = \mathrm{Crop}(\mathcal{V}_I, b^*).$$

This removes irrelevant background while preserving the relative positions and spatial layout required for subsequent relation reasoning.

4.5 Evidence Tracing and Answering

After Spatial Intent Routing, the agent performs Image Space Tracing and Object Space Tracing as needed to obtain local evidence. Let the final visual input be $\mathcal{V}^*$; the prediction is $\hat{y} = F_\theta(\mathcal{V}^*, Q)$, where F_θ is the frozen base VLM. For image position questions, the answer is determined directly from candidate verification scores; otherwise, the VLM predicts from the traced evidence and original question.

Table 2: Comparison with existing methods on the XLRS-Bench Lite benchmark. The best and second-best results are highlighted in **bold** and underlined, respectively.

Method	Perception				Reasoning				Avg.
	Cnt	SC	OSR	OP	Plan	AR	CR	SR	
GPT-4o	28.75	40.33	24.60	10.78	35.00	73.00	49.00	20.00	22.30
Claude-3.7-Sonnet	25.00	43.00	27.60	41.33	35.00	65.00	55.50	28.33	39.65
LLaVA-v1.5-7B	22.50	19.67	23.00	20.18	39.00	35.00	34.00	30.00	22.89
Qwen2.5-VL-7B	30.63	40.33	30.80	43.31	32.00	67.00	62.50	43.33	41.98
LLaVA-OV-7B	31.25	35.67	25.20	45.42	24.00	76.00	59.50	43.33	41.62
LLaVA-NeXT	33.13	36.00	30.00	42.71	27.00	69.00	57.00	35.00	40.62
GeoChat	22.50	12.67	24.20	24.28	10.00	33.00	32.00	–	23.35
ZoomEarth	–	–	–	–	–	–	–	–	40.20
GeoLLaVA-8K	35.00	62.00	35.80	37.71	**65.00**	68.00	63.50	51.67	43.44
C2F	35.00	35.33	26.80	33.37	24.00	69.00	76.00	**58.33**	36.69
ZoomSearch	36.25	48.33	28.80	39.52	33.00	76.00	**81.00**	45.00	42.24
GeoEyes	39.36	57.00	32.20	63.00	56.00	**77.00**	65.00	40.00	56.10
GeoVista	41.25	54.00	**42.40**	52.47	46.00	69.00	52.50	48.33	50.65
WeaveEarth	37.50	50.33	36.40	47.41	36.00	73.00	67.50	46.33	47.14
TraceEarth (Ours)	**48.75**	**66.33**	31.60	**70.72**	37.00	74.00	67.50	45.00	**61.10**

5 EXPERIMENT

5.1 EXPERIMENTAL SETUP

Datasets. We evaluate TraceEarth on three remote sensing VQA benchmarks: MME-RealWorld-RS Zhang et al. (2024b), XLRS-Bench Wang et al. (2025b), and LHRS-Bench Muhtar et al. (2025). MME-RealWorld-RS covers Color, Count, and Position, XLRS-Bench evaluates diverse UHR perception and reasoning tasks, and LHRS-Bench provides additional evaluation across general remote sensing scenarios. We report accuracy for each task and overall accuracy weighted by sample size on all three benchmarks.

Implementation Details. TraceEarth is fully training free and performs no parameter updates during inference. We use Qwen3-VL-8B-Instruct as the primary base VLM, and evaluate LLaVA-OV-Qwen2-7B and Qwen3-VL-32B for generalization across model scales and architectures. The frozen base model executes spatial intent routing, image space tracing, and object space tracing with stage specific instructions and spatial skills, while RemoteSAM (Yao et al., 2025) is used for reference entity localization. All experiments are conducted on NVIDIA RTX A6000 GPUs. Structured stage inputs and outputs, along with skill implementations, are provided in Appendix A and Appendix B, respectively.

5.2 BASELINES.

We compare TraceEarth with representative remote sensing and UHR perception methods, including GeoChat (Kuckreja et al., 2024), VHM-7B (Pang et al., 2024), GeoLLaVA-8K (Wang et al., 2025a), GeoVista (Zhu et al., 2026), ZoomEye (Shen et al., 2024), ImageRAG (Zhang et al., 2025c), RAP (Wang et al.), RADAR (Liu et al., 2026b), C2F (Luo et al., 2025), ZoomSearch (Zhou et al., 2025), ZoomEarth (Liu et al., 2026a), GeoEyes (Wang et al., 2026a), and WeaveEarth (Ma et al., 2026b), using available results under their original evaluation settings.

5.3 MAIN RESULTS

Tables 1 and 2 report the main results on MME-RealWorld-RS and XLRS-Bench Lite. TraceEarth achieves overall accuracies of 60.03% and 61.10%, improving over the strongest compared methods by 2.39 and 5.00 percentage points, respectively. It performs particularly well on spatial localization tasks, reaching 77.72% on MME-RealWorld-RS Position and 70.72% on XLRS-Bench Lite OP. These results show that converting spatial cues into tracing paths supports visual evidence localization in UHR scenes. TraceEarth further achieves 72.61% on LHRS-Bench, indicating that spatial tracing also generalizes beyond UHR imagery to remote sensing VQA at different image scales.

Table 3: Ablation study on MME-RealWorld-RS. SIR, IST, and OST denote spatial intent routing, image space tracing, and object space tracing, respectively.

Method	SIR	IST	OST	Color	Count	Position	Avg.
Qwen3-VL-8B	–	–	–	59.76	21.85	68.01	50.10
w/o SIR & IST	–	–	✓	63.57	23.89	73.48	53.91
w/o SIR	–	✓	✓	67.25	28.30	73.74	56.66
TraceEarth (Ours)	✓	✓	✓	**69.32**	**32.38**	**77.72**	**60.03**

Table 4: Generalization of TraceEarth across VLM backbones on MME-RealWorld-RS.

Backbone	Variant	Color	Count	Position	Avg.
Qwen3-VL-4B	Vanilla	55.05	13.21	62.45	43.82
	+ TraceEarth	61.51 (+6.46)	16.88 (+3.67)	76.13 (+13.68)	51.79 (+7.97)
LLaVA-OV-7B	Vanilla	55.93	22.51	63.80	47.62
	+ TraceEarth	63.19 (+7.26)	37.27 (+14.76)	73.19 (+9.39)	58.05 (+10.43)
Qwen3-VL-8B	Vanilla	59.76	21.85	68.01	50.10
	+ TraceEarth	69.32 (+9.56)	32.38 (+10.53)	77.72 (+9.71)	60.03 (+9.93)
Qwen3-VL-32B	Vanilla	57.93	23.57	72.08	51.42
	+ TraceEarth	68.69 (+10.76)	31.32 (+7.75)	77.41 (+5.33)	59.36 (+7.94)

5.4 Ablation Study

We evaluate the contribution of each component on MME-RealWorld-RS by progressively introducing object space tracing, image space tracing, and Spatial Intent Routing under the same base VLM and evaluation settings. As shown in Table 3, the vanilla Qwen3-VL-8B achieves 50.10% overall accuracy. Introducing Object Space Tracing improves the accuracy to 53.91%, while further adding image space tracing raises it to 56.66%. With Spatial Intent Routing enabled, the full TraceEarth reaches 60.03%. These consistent gains show that Object Space and image space provide complementary spatial constraints for local evidence acquisition, while Spatial Intent Routing further improves performance by selecting appropriate tracing strategies according to the role of spatial cues.

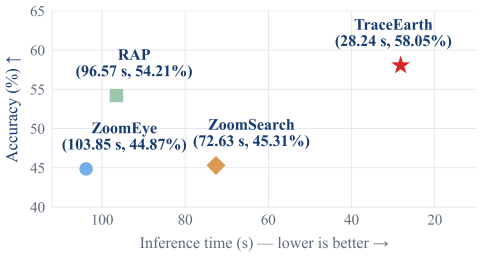


Figure 4: Accuracy and inference efficiency comparison on MME-RealWorld-RS.

5.5 Generalization across VLM Backbones

We evaluate TraceEarth across VLMs with different architectures and scales, including Qwen3-VL-4B, LLaVA-OneVision-7B, Qwen3-VL-8B, and Qwen3-VL-32B. As shown in Table 4, TraceEarth consistently improves overall accuracy by 7.97, 10.43, 9.93, and 7.94 percentage points, respectively. These consistent gains show that TraceEarth is not tied to a specific model architecture or scale, supporting its general applicability as a training-free visual evidence acquisition framework.

5.6 Accuracy–Efficiency Comparison

To evaluate inference efficiency, we compare TraceEarth with ZoomEye, RAP, and ZoomSearch using the same LLaVA-OV-7B backbone. As shown in Figure 4, TraceEarth achieves the highest accuracy of 58.05% and the lowest inference time of 28.24 s among the compared methods. These results demonstrate that spatial tracing enables efficient visual evidence acquisition while maintaining strong VQA performance.

Table 5: Plug-and-play evaluation of TraceEarth on RAP on MME-RealWorld. We report accuracy and inference time for the Color, Count, and Position subsets.

Method	Accuracy (%) ↑				Time (s) ↓			
	Color	Count	Position	Avg.	Color	Count	Position	Avg.
RAP	64.18	39.15	58.95	54.21	98.12	68.82	122.77	96.57
RAP + TraceEarth	66.53	39.72	63.07	56.75	60.22	56.80	94.76	70.59
Δ	*+2.35*	*+0.57*	*+4.12*	*+2.54*	*-37.90*	*-12.02*	*-28.01*	*-25.98*

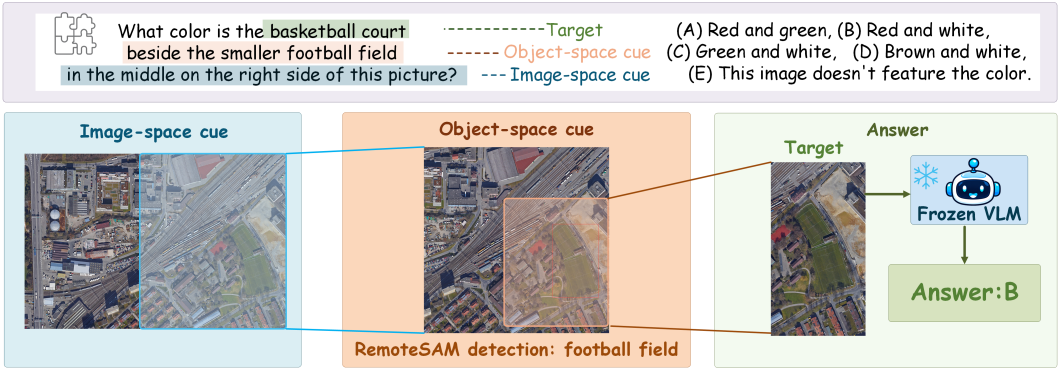


Figure 5: Case study of TraceEarth. Image Space tracing first narrows the search region using the image-level cue, after which RemoteSAM localizes the reference football field and Object Space tracing follows the spatial relation to reach the target basketball court. The traced local evidence is then provided to the frozen VLM for final answer prediction.

5.7 Plug-and-Play Spatial Tracing

To evaluate the compatibility of TraceEarth with existing UHR visual search methods, we integrate it into RAP while keeping its base model and overall inference pipeline unchanged. As shown in Table 5, TraceEarth increases the average accuracy from 54.21% to 56.75% (+2.54 points) while reducing inference time from 96.57 s to 70.59 s (-25.98 s). Both accuracy and efficiency improve consistently across all three subsets, demonstrating that TraceEarth can serve as an effective plug and play module for existing visual search methods.

5.8 Case Study and Discussion

To illustrate the spatial tracing process of TraceEarth, we present a representative case in Fig. 5. TraceEarth first uses the Image Space cue to narrow the search region, then localizes the football field with RemoteSAM and follows the Object Space relation to trace the target basketball court. The resulting local evidence is finally provided to the frozen VLM for answer prediction. This demonstrates how TraceEarth progressively follows spatial dependencies from the global scene to the local target while reducing irrelevant visual content.

6 Conclusion

We present TraceEarth, a training-free framework for spatially guided visual evidence acquisition in UHR remote sensing VQA. TraceEarth decouples spatial cues from the full question semantics and organizes them in image space and object space, enabling visual evidence to be progressively traced from global scenes to local targets and shifting evidence acquisition from semantic matching to spatial tracing. Experiments across multiple benchmarks and VLM backbones, together with plug and play integration, demonstrate its effectiveness, generalization, and compatibility with existing visual search methods.

AI USE STATEMENT

Generative AI tools were used to assist with language editing, translation, and improving the clarity of the manuscript, as well as for limited assistance with code debugging and implementation. All research ideas, methodological designs, experimental settings, results, and scientific conclusions were developed, conducted, and verified by the authors. All AI-assisted text and code were manually reviewed and checked by the authors.

REFERENCES

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

Xin Guo, Jiangwei Lao, Bo Dang, Yingying Zhang, Lei Yu, Lixiang Ru, Liheng Zhong, Ziyuan Huang, Kang Wu, Dingxiang Hu, Huimei He, Jian Wang, Jingdong Chen, Ming Yang, Yongjun Zhang, and Yansheng Li. Skysense: A multi-modal remote sensing foundation model towards universal interpretation for earth observation imagery. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 27672–27683, June 2024.

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.

Yao kelu, Xu Nuo, Yang Rong, Xu Yingying, Gao Zhuoyan, Kitrungrotsakul Titinunt, Ren yi, Zhang Pu, Wang Jin, Wei Ning, and Li Chao. Falcon: A remote sensing vision-language foundation model. *arXiv preprint arXiv:2503.11070*, 2025.

Kartik Kuckreja, Muhammad S. Danish, Muzammal Naseer, Abhijit Das, Salman Khan, and Fahad S. Khan. Geochat: Grounded large vision-language model for remote sensing. *The IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

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

Geng Li, Jinglin Xu, Yunzhen Zhao, and Yuxin Peng. Dyfo: A training-free dynamic focus visual search for enhancing lmms in fine-grained visual understanding, 2025. URL https://arxiv.org/abs/2504.14920.

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. *arXiv preprint arXiv:2603.03857*, 2026b.

Yueying Li, Fengxiang Wang, Yan Li, Mingshuo Chen, Mengying Zhao, and Long Lan. Semantic-geometric dual compression: Training-free visual token reduction for ultra-high-resolution remote sensing understanding, 2026c. URL https://arxiv.org/abs/2604.11122.

Zhenshi Li, Dilxat Muhtar, Feng Gu, Xueliang Zhang, Pengfeng Xiao, Guangjun He, and Xiaoxiang Zhu. Lhrs-bot-nova: Improved multimodal large language model for remote sensing vision-language interpretation. *arXiv preprint arXiv:2411.09301*, 2024.

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

Wang Liu, Puhong Duan, Xudong Kang, and Shutao Li. Squeezing context into patches: Towards memory-efficient ultra-high resolution semantic segmentation. In James Kwok (ed.), *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence, IJCAI-25*, pp. 1603–1611. International Joint Conferences on Artificial Intelligence Organization, 8 2025. doi:

10.24963/ijcai.2025/179. URL https://doi.org/10.24963/ijcai.2025/179. Main Track.

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, 2026b. 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 (ICCV)*, pp. 9206–9217, October 2025.

Kexin Ma, Jing Xiao, Bowen Xing, Liang Liao, and Chia-Wen Lin. Sa-gem: Scale-adaptive and geospatial evidence-modulated token pruning for efficient remote sensing large vision-language models, 2026a. URL https://arxiv.org/abs/2608.15075.

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, 2026b. 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, Cham, 2025. Springer Nature Switzerland. ISBN 978-3-031-72904-1. doi: 10.1007/978-3-031-72904-1_26.

Chao Pang, Xingxing Weng, Jiang Wu, Jiayu Li, Yi Liu, Jiaxing Sun, Weijia Li, Shuai Wang, Litong Feng, Gui-Song Xia, and Conghui He. Vhm: Versatile and honest vision language model for remote sensing image analysis, 2024. URL https://arxiv.org/abs/2403.20213.

Ji Qi, Ming Ding, Weihan Wang, Yushi Bai, Qingsong Lv, Wenyi Hong, Bin Xu, Lei Hou, Juanzi Li, Yuxiao Dong, and Jie Tang. Cogcom: Train large vision-language models diving into details through chain of manipulations. *arXiv preprint arXiv:2402.04236*, 2024.

Hiyam Debary Mustansar Fiaz Muhammad Akhtar Munir Muhammad Sohail Danisho Paolo Fraccaro Campbell Watson Levente J. Klein Salman Khan Fahad Khan Sagar Soni, Akshay Dudhane. Earthdial: Turning multi-sensory earth observations to interactive dialogues. *ArXiv*, 2025. URL https://arxiv.org/abs/2412.15190.

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. *arXiv preprint arXiv:2411.16044*, 2024.

Alex Su, Haozhe Wang, Weiming Ren, Fangzhen Lin, and Wenhu Chen. Pixel reasoner: Incentivizing pixel-space reasoning with curiosity-driven reinforcement learning. *arXiv preprint arXiv:2505.15966*, 2025.

Fengxiang Wang, Mingshuo Chen, Yueying Li, Di Wang, Haotian Wang, Zonghao Guo, Zefan Wang, Boqi Shan, Long Lan, Yulin Wang, Hongzhen Wang, Wenjing Yang, Bo Du, and Jing Zhang. Geollava-8k: Scaling remote-sensing multimodal large language models to 8k resolution. *arXiv preprint arXiv:2505.21375*, 2025a.

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

Fengxiang Wang, Mingshuo Chen, Yueying Li, Yajie Yang, Yifan Zhang, Long Lan, Xue Yang, Hongda Sun, Yulin Wang, Di Wang, et al. Geoeyes: On-demand visual focusing for evidence-grounded understanding of ultra-high-resolution remote sensing imagery. *arXiv preprint arXiv:2602.14201*, 2026a.

Fengxiang Wang, Mingshuo Chen, Yueying Li, Yajie Yang, Yuhao Zhou, Di Wang, Yifan Zhang, Haoyu Wang, Haiyan Zhao, Hongda Sun, et al. Text before vision: Staged knowledge injection matters for agentic rlvr in ultra-high-resolution remote sensing understanding. *arXiv preprint arXiv:2602.14225*, 2026b.

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 *Forty-second International Conference on Machine Learning*. URL https://arxiv.org/abs/2503.01222.

Kang Wu, Yingying Zhang, Lixiang Ru, Bo Dang, Jiangwei Lao, Lei Yu, Junwei Luo, Zifan Zhu, Yue Sun, Jiahao Zhang, Qi Zhu, Jian Wang, Ming Yang, Jingdong Chen, Yongjun Zhang, and Yansheng Li. A semantic-enhanced multi-modal remote sensing foundation model for earth observation. *Nature Machine Intelligence*, 2025. doi: 10.1038/s42256-025-01078-8. URL https://doi.org/10.1038/s42256-025-01078-8.

Penghao Wu and Saining Xie. V*: Guided visual search as a core mechanism in multimodal llms. *arXiv preprint arXiv:2312.14135*, 2023.

Liang Yao, Fan Liu, Delong Chen, Chuanyi Zhang, Yijun Wang, Ziyun Chen, Wei Xu, Shimin Di, and Yuhui Zheng. Remotesam: Towards segment anything for earth observation, 2025. URL https://arxiv.org/abs/2505.18022.

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

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 *The Thirteenth International Conference on Learning Representations*, 2025a. URL https://arxiv.org/abs/2502.17422.

Peirong Zhang, Yidan Zhang, Luxiao Xu, Jinliang Lin, Zonghao Guo, Fengxiang Wang, Xue Yang, Kaiwen Wei, and Lei Wang. Geovis: Geospatially rewarded visual search for remote sensing visual grounding, 2025b. URL https://arxiv.org/abs/2512.02715.

Wei Zhang, Miaoxin Cai, Tong Zhang, Yin Zhuang, and Xuerui Mao. Earthgpt: A universal multimodal large language model for multi-sensor image comprehension in remote sensing domain. *IEEE Transactions on Geoscience and Remote Sensing*, 2024a.

Yi-Fan Zhang, Huanyu Zhang, Haochen Tian, Chaoyou Fu, Shuangqing Zhang, Junfei Wu, Feng Li, Kun Wang, Qingsong Wen, Zhang Zhang, et al. Mme-realworld: Could your multimodal llm challenge high-resolution real-world scenarios that are difficult for humans? *arXiv preprint arXiv:2408.13257*, 2024b.

Zilun Zhang, Haozhan Shen, Tiancheng Zhao, Zian Guan, Bin Chen, Yuhao Wang, Xu Jia, Yuxiang Cai, Yongheng Shang, and Jianwei Yin. Enhancing ultrahigh resolution remote sensing imagery analysis with imagerag: A new framework. *IEEE Geoscience and Remote Sensing Magazine*, pp. 2–27, 2025c. doi: 10.1109/MGRS.2025.3574742.

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. *arXiv preprint arXiv:2505.14362*, 2025.

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.

Jiashun Zhu, Ronghao Fu, Jiasen Hu, Jing Huang, Nachuan Xing, and Bo Yang. Geovista: Visually grounded active perception for vision-language understanding of ultra-high-resolution remote sensing images. *arXiv preprint arXiv:2605.14475*, 2026.

A TRACEEARTH AGENT PROMPTS AND STRUCTURED OUTPUTS

TraceEarth employs a unified spatial tracing Agent and uses stage specific task instructions to perform different types of spatial parsing and tracing decisions. Throughout the process, all stages share the same frozen base vision language model, while the prompts and structured output formats are switched according to the current task state. Specifically, the TraceEarth Agent sequentially performs Spatial Intent Routing, Image Space parsing, and Object Space parsing, progressively transforming spatial information in the question into structured constraints for subsequent visual tracing. Table 6 summarizes the main function and output format of each stage.

Table 6: Overview of the prompt functions and structured outputs used at different stages of the TraceEarth Agent.

Stage	**Main Function**	**Structured Output**
Spatial Intent Routing	Identify the role of spatial information in the current question and determine the subsequent spatial tracing and evidence acquisition path.	Routing result, together with the target entity, spatial reference, and linguistic evidence relevant to the current spatial decision.
Image Space Parsing	Parse spatial information referenced to the global image coordinate system and determine the corresponding executable spatial operation.	Image Space reference type, spatial parameters, target object or scope, and the recommended spatial Skill.
Object Space Parsing	Parse the target entity, reference entity, and their spatial dependencies to provide structured constraints for relation guided tracing or relation preserving evidence construction.	Target and reference entities, known or unknown spatial relations, local spatial constraints, and candidate entities available for localization.

A.1 SPATIAL INTENT ROUTING AND STRUCTURED OUTPUT

Spatial Intent Routing identifies the role of spatial information in the current question and determines how visual evidence should subsequently be acquired. The TraceEarth Agent produces a unified structured output {`route, target, anchor, evidence`}, where `route` specifies how the spatial information should be used, `target` and `anchor` denote the prediction target and its spatial reference, respectively, and `evidence` preserves the original question span supporting the routing decision. The routing types and their corresponding spatial tracing strategies are summarized in Table 7.

Table 7: Spatial Intent Routing and the corresponding use of spatial information.

Route	**Role of Spatial Information**	**Subsequent Processing**
`image_direction`	The location to be predicted is referenced to the entire image.	Convert candidate locations into spatial tracing paths to be verified in Image Space and determine the location based on local visual evidence.
`object_direction`	The location to be predicted is referenced to a scene entity.	Preserve the relevant entities and their spatial layout, leaving the unknown relation to be determined from subsequent visual evidence.
`general`	Spatial information serves as a known localization condition rather than the final prediction target.	Use the known spatial constraints to perform Image Space Tracing and Object Space Tracing as needed for local visual evidence acquisition.

A.2 Image Space Parsing and Structured Output

For questions containing Image Space information, the TraceEarth Agent further parses spatial cues referenced to the current image coordinate system and represents them as structured Image Space references. This stage mainly identifies three types of spatial representations: image regions described in natural language, explicit bounding boxes, and visually marked regions, and matches each representation to the corresponding spatial Skill. This stage only performs spatial parsing and Skill matching, while the actual visual operations are executed during the subsequent tracing process. The different Image Space references and their corresponding structured representations are summarized in Table 8.

Table 8: Main structured outputs of the Image Space parsing stage.

Structured Information	**Meaning**	**Subsequent Use**
`grid`	A natural language region description referenced to the entire image, together with the parsed region direction or image grid location.	Invoke `grid_crop` to extract the corresponding image region.
`bbox`	Explicit bounding box information provided in the question, including coordinate values, format, and units.	Invoke `bbox_crop` to extract the specified region.
`visual_marker`	An image region specified by color, shape, or drawn markers, together with the parsed visual marker attributes.	Invoke `marker_localize_crop` to localize and extract the marked region.

For each parsed Image Space reference, the Agent also preserves its original linguistic evidence, target object or scope, spatial dependencies, and any uncertainty. This avoids introducing localization information that is not explicitly provided in the question during spatial parsing.

Candidate Conditioned Question Transformation. For questions where the image position itself is the prediction target, candidate locations cannot be directly used as known spatial constraints. Therefore, once Spatial Intent Routing assigns the question to `image_direction`, the TraceEarth Agent further invokes a candidate transformation instruction to convert the original multiple choice position question into several independent local evidence verification branches.

Specifically, the Agent first extracts the visual target to be localized from the question. It then preserves the original label and text of each candidate answer and normalizes its spatial description into a canonical Image Space region. For candidate locations that can be mapped to the image grid, the system further identifies the corresponding 3×3 grid cells and generates a local binary verification question for that candidate region. In this way, each candidate answer is transformed into a structured representation of "candidate spatial path–local region–evidence verification," which is subsequently used by Image Space Tracing for region cropping and visual verification.

A.3 Object Space Parsing and Structured Output

For questions containing Object Space information, the TraceEarth Agent further parses the target entity, reference entity, and their spatial dependencies, and organizes them into a structured representation for subsequent Object Space Tracing. Unlike Image Space constraints that can be directly mapped to image regions, localization in Object Space requires progressively resolving scene entities and their relations. This stage therefore distinguishes known spatial relations from relations to be inferred, while extracting candidate references that can support subsequent entity localization and relation guided tracing. The main structured outputs are summarized in Table 10.

A.4 Integration of Structured Outputs with Spatial Tracing

The three stages above jointly form a structured interface between question understanding and visual execution. Spatial Intent Routing first determines how spatial information should be used in the

Table 9: Main structured outputs of the candidate conditioned question transformation stage.

Structured Information	Meaning	Subsequent Use
`target`	The visual target to be localized, extracted from the original question.	Serves as the common query target for local verification across all candidate regions.
`canonical_region`	The canonical image region normalized from the spatial description in a candidate answer.	Provides the standardized representation of the candidate location in Image Space.
`grid_cells`	The 3×3 image grid cells corresponding to the canonical region.	Used to construct the candidate region and perform subsequent local cropping.
`verification_question`	A local Yes/No verification question generated for the current candidate region.	Combined with the candidate local image as input to the VLM to compute an evidence support score for the spatial hypothesis.
`absence_option`	A candidate answer indicating that the target is absent or not visible.	No local cropping branch is generated, and the option is retained for final position prediction.

Table 10: Main structured outputs of the Object Space parsing stage.

Structured Information	Meaning	Subsequent Use
`target / references`	The target entity in the question and the reference entities that define its spatial location.	Provides entity information for reference entity localization and subsequent relation guided tracing.
`relations`	Spatial relations between objects that are explicitly given in the question.	Used as known spatial constraints to guide Object Space Tracing toward the target region.
`local_constraints`	Local directional constraints within the target or reference entity.	Further narrows the visual search range within the entity or around its boundary.
`selections`	Selection conditions used to identify a specific instance among multiple entities of the same category.	Helps determine the specific reference instance used for subsequent spatial tracing.
`relation_queries`	Unknown spatial relations in the current question that must be inferred from visual evidence.	Not used as tracing constraints in advance; instead, the relevant entities and their spatial layout are preserved during evidence acquisition.
`localization_candidates`	Candidate entities extracted from the question text that can help narrow the visual search range.	Provides candidate objects for reference entity grounding and subsequent spatial Skill invocation.

current question. Image Space Parsing and Object Space Parsing then extract image level spatial constraints and object level spatial dependencies that can be used for visual localization, respectively. These structured outputs further determine how subsequent visual evidence should be acquired, as summarized in Table 11.

Table 11: Correspondence between Agent structured outputs and subsequent spatial tracing.

Stage	**Key Output**	**Role in Subsequent Tracing**
Spatial Intent Routing	`route`, `target`, `anchor`	Determine how spatial information should be used and select the corresponding visual evidence acquisition and verification path.
Image Space Parsing	Image reference type, spatial parameters, and recommended Skill	Map spatial cues referenced to the entire image into executable region localization or cropping operations.
Object Space Parsing	Reference entities, known spatial relations, local constraints, and relations to be inferred	Use known relations to continue Object Space Tracing; for relations to be inferred, preserve the relevant entities and their spatial layout for final prediction.

In this way, the TraceEarth Agent connects structured spatial information parsed at the language level to actual visual execution. Known spatial constraints can be directly used to narrow the visual search space, while spatial relations that remain to be predicted are not used as tracing constraints in advance, thereby avoiding the introduction of unknown information during evidence acquisition.

B Spatial Skill Library and Execution Details

B.1 Spatial Skill Library and Unified Execution Interface

To translate the structured spatial constraints parsed by the Agent into actual visual operations, TraceEarth builds a set of lightweight spatial Skills as a unified execution interface between language level spatial parsing and visual evidence acquisition. Different Skills handle image region constraints in Image Space and entity localization or relation guided tracing in Object Space. The TraceEarth Agent selects the corresponding Skill according to the structured output of the current stage and generates the spatial parameters required for execution. The main Skills are summarized in Table 12.

These Skills do not answer the question themselves, but only execute the spatial operations determined by the TraceEarth Agent. Spatial parsing, Skill selection, and visual execution are therefore separated, allowing the same tracing framework to handle different forms of spatial references and consistently convert them into local visual evidence.

B.2 Execution Details of Image Space Skills

This section provides implementation details of the spatial Skills and candidate verification process used in Image Space Tracing.

`grid_crop`. For natural language directional descriptions, TraceEarth divides the image into a regular 3×3 grid. A single region, such as `upper-left`, is directly mapped to the corresponding grid cell, while broader descriptions, such as `left` or `upper`, are cropped using the enclosing region of the relevant grid cells. The region mapping rules are consistent with those used for candidate location transformation.

`bbox_crop`. For explicit bounding boxes, the parsed coordinates are first converted to the current image coordinate system according to their format and units, and the corresponding region is then cropped. Coordinates outside the image range are clipped to valid image boundaries.

`marker_localize_crop`. For regions specified by color, shape, or drawn markers, the corresponding visual marker is first localized. A crop is then generated according to its spatial extent while preserving local visual context around the marked region.

Table 12: Main spatial Skills used in TraceEarth.

Space	Spatial Skill	Function
Image Space	`grid_crop`	Map and crop the full image to the corresponding region according to a natural language image direction or canonical grid location.
Image Space	`bbox_crop`	Extract the corresponding local visual region according to the bounding box coordinates explicitly provided in the question.
Image Space	`marker_localize_crop`	Localize a target region according to visual cues such as color, shape, or drawn markers, and extract the corresponding local image.
Object Space	`entity_localize`	Localize the specified scene entity within the current visual region and obtain its spatial position, providing a visual anchor for subsequent Object Space Tracing.
Object Space	`relation_guided_trace`	Use the localized entity as an anchor and expand the visual search region along the direction indicated by a known spatial relation in the question to acquire target related evidence.
Object Space	`relation_preserving_crop`	When the spatial relation itself is to be inferred, jointly localize the relevant entities and preserve their original spatial layout to construct a compact visual region containing relation evidence.

Candidate Verification. For the `image_direction` route, the structured transformation of candidate questions is performed by the Agent described in Appendix A.2. For each candidate region V_i and its verification question $\widetilde{Q}_i$, we obtain the logits corresponding to `Yes` and `No` at the next output position of the frozen VLM and compute the verification support score:

$$s_i = \frac{\exp(z_i^+)}{\exp(z_i^+) + \exp(z_i^-)}. \tag{1}$$

If the gap between the highest and second highest scores is smaller than $\tau = 0.03$, the full image and original question are used for fallback inference. Otherwise, the candidate with the highest verification support score is selected.

B.3 Execution Details of Object Space Skills

This section provides implementation details of the spatial Skills used in Object Space Tracing. The parsing and structured outputs of Object Space information are described in Appendix A.3.

`entity_localize`. For questions requiring entity grounding, TraceEarth uses RemoteSAM to localize the specified entity within the current visual region and obtain its bounding box set:

$$\mathcal{B}_o = D(V, o), \tag{2}$$

where o denotes the entity to be localized and $D(\cdot)$ denotes the entity localizer. When multiple matching instances are detected, the instance selection conditions parsed by the Agent, such as local direction, quantity, or attribute constraints, are used to determine the entity location for subsequent tracing.

`relation_guided_trace`. When a known object spatial relation is provided in the question, the localized reference entity is used as the visual anchor and the corresponding search region is constructed according to the relation direction. For relations with a clear dominant direction, the search region extends along that direction to the corresponding boundary of the current visual region. Meanwhile, context equal to 0.5 times the reference entity scale is preserved in the opposite and

orthogonal directions to fully retain the reference entity and avoid missing nearby targets. The resulting relation guided local visual region is:

$$V_O = G_O(V, \mathcal{B}_a, \mathcal{R}_k), \tag{3}$$

where $\mathcal{B}_a$ denotes the reference entity location and $\mathcal{R}_k$ denotes the known spatial relation.

`relation_preserving_crop`. When the spatial relation itself is the prediction target, it is not used to determine the cropping direction. TraceEarth first localizes the entities relevant to relation reasoning and obtains the bounding box set:

$$\mathcal{B} = \{b_1, \ldots, b_n\}, \tag{4}$$

then computes the smallest enclosing region covering these entities and constructs the local visual input while preserving their original relative layout:

$$b^* = \mathrm{Envelope}(\mathcal{B}), \qquad V_O = \mathrm{Crop}(V, b^*). \tag{5}$$

This operation only removes irrelevant background without altering the spatial relations among the relevant entities, allowing the subsequent VLM to directly infer the relation from local visual evidence.