

Asking like Socrates: Socrates helps VLMs understand remote sensing images

Supplementary Material

1. System Prompt Details

In this section, we provide the exact system prompts used in our SocraticAgent framework to synthesize the RS-EoT-4K dataset. As described in the main paper, SocraticAgent operates as a self-play multi-agent system consisting of three distinct roles: the **Reasoner**, the **Perceiver**, and the **Verifier**.

- **The Reasoner** (Fig. 1) serves as the cognitive core. It is a text-only agent designed to simulate the human “coarse-to-fine” visual interpretation process. By following a *Plan–Integrate–Decide* paradigm, it decomposes complex queries into atomic visual questions, iteratively gathering evidence before concluding.
- **The Perceiver** (Fig. 2) acts as the visual interface. To enforce explicit and detailed reasoning from the Reasoner, we employ a self-play prompting mechanism where the Perceiver is instructed to collaborate with a “weak-reasoning” partner. Its role is to provide accurate, descriptive visual observations in a “natural inner monologue” style without performing high-level logical reasoning.
- **The Verifier** (Fig. 3) functions as the quality control gatekeeper. It compares the final answer generated by the Reasoner against the ground truth to filter out incorrect or hallucinated reasoning traces during the data synthesis phase.

The complete prompts for these agents are presented in Fig. 1, Fig. 2, and Fig. 3, respectively.

2. SFT Training Settings

We perform SFT on the base model Qwen2.5-VL-7B-Instruct using the RS-EoT-4K dataset. The training is implemented based on the LLaMA-Factory framework. We train the model for 5 epochs with a learning rate of 3×10^{-5} , using the AdamW optimizer and a cosine learning rate scheduler. The global batch size is set to 64, and the maximum sequence length is limited to 4096 tokens to accommodate the detailed reasoning traces. The entire training process was executed on 4 A100 GPUs and required approximately 40 minutes to complete.

An empirical finding in our experiments concerns the use of system prompts. We observed that including explicit reasoning triggers in the system prompt (e.g., instructions like “Please reason step-by-step”) during training caused the model to develop a strong dependency on these specific triggers. Consequently, the model often failed to initiate reasoning or exhibited abnormal behaviors when such prompts were absent during inference. To address this and ensure robust, spontaneous reasoning capabilities, we exclude all

reasoning-specific instructions from the system prompt during SFT and RL. Instead, we hard-code the start-of-thought token `<think>` directly into the assistant’s response field within the chat template. This design forces the model to automatically enter the reasoning mode immediately upon generation, ensuring stable output trajectories without relying on specific user or system prompts.

Training and evaluation data usage. The synthetic trajectories generated by SocraticAgent are used exclusively for training. Concretely, our pipeline consists of three stages: (1) collecting queries from public remote-sensing datasets, (2) synthesizing RS-EoT trajectories for SFT, and (3) performing the subsequent two-stage RL procedure. All results reported in the main paper and this supplementary material are evaluated on public benchmarks with their official splits; no synthetic data are used for evaluation.

Grounding degradation after SFT. We observe a temporary drop in grounding performance immediately after SFT because the SFT corpus contains VQA-style reasoning trajectories but no grounding samples. In our pipeline, the role of SFT is to initialize the reasoning format and stabilize long-form thought generation rather than to maximize grounding accuracy directly. The subsequent RL-Grounding stage recovers and further improves the grounding ability once grounding-specific rewards are introduced.

3. RL Training Settings

All reinforcement learning experiments are conducted using the EasyR1 framework, which provides a production-ready implementation of GRPO with KL regularization. We fix the KL coefficient to $\beta = 1.0 \times 10^{-2}$. For each input, the model generates 4 rollout samples using sampling temperature 1.0, with a maximum response length of 4096 tokens.

Policy optimization is performed with the AdamW optimizer, using a learning rate of 1.0×10^{-6} , weight decay of 1.0×10^{-2} , and a linear warm-up over the first 3% of training steps. We apply gradient clipping with a threshold of 1.0. All models are trained in `bf16` precision under full-sharded data parallelism (FSDP) with gradient checkpointing enabled.

For both RL-Grounding and RL-VQA, we use a global batch size of 512 (covering rollout, training, and validation batches) and train for 240 steps. Experiments are run on a single node equipped with 8 A100 GPUs. The RL-VQA stage requires approximately 2.3 days to complete, while the RL-Grounding stage takes about 2.5 days.

System Prompt for the Reasoner

You are a reasoning model that follows a **Plan–Integrate–Decide** paradigm, collaborating with a **weak-perception visual model** to complete **general remote sensing tasks** (such as classification/attribute recognition, localization/counting, relation/change detection, and VQA). The perception model can only answer **very simple, atomic visual facts** and cannot perform reasoning. Therefore, you must decompose the perception process into a **coarse-to-fine sequence of steps**, simulating how humans visually interpret remote sensing imagery.

[Coarse-to-Fine Perception Chain]

1) Global Observation Stage (Overall Understanding):

- Begin with a **broad, holistic examination** of the entire image, forming an initial impression of its overall layout — main land-cover types, spatial organization, scene functionality, distribution of major objects, and possible visual interferences (e.g., shadows, fog, noise, or occlusion).
- While questions at this stage should remain **broad, general, and high-level**, they must be **context-aware** — i.e., lightly tailored to the task/query so they inform later reasoning for this specific problem.

2) Focused and Detailed Observation Stage (Targeted Analysis):

- After forming a general understanding of the scene, use the task objective (query) and global observations to **focus attention** on potentially relevant local regions or objects.
- Naturally shift attention from overall impressions to specific, task-relevant areas, similar to how humans visually focus.
- Ask more detailed and targeted questions, typically focusing on:
 - Local details (shape, texture, boundaries, orientation, color features, etc.);
 - Relationships and differences (changes, similarities, transitions between regions, etc.);
 - Task-critical elements (e.g., presence, quantity, or arrangement of specific targets).

3) Integration and Verification Stage:

- Integrate the facts collected from the global and detailed observation stages into a consistent intermediate conclusion.
- If contradictions or uncertainties remain, ask verification questions.
- Ensure that the reasoning covers all key regions and that the logic is consistent.

4) Final Review and Confirmation Stage:

- Before giving the final answer, perform a quick overall review of the image to confirm that no small anomalies, marginal areas, or potential clues have been overlooked.
- Check whether the final answer meets the query’s requirements regarding format, length, and structure.
- The final answer must only output the direct answer to the query itself, such as “Yes/No”, a specific number, or a concise conclusion. Do not include any explanations, reasoning, or additional commentary.
- If necessary, ask one final targeted question for confirmation.

[Questioning and Iteration Constraints]

- **Never** forward the user’s original query directly to the perception model; each question must concern **only one atomic visual fact**.
- Each new question should provide **maximum information gain** and **must not repeat** previous questions (avoid paraphrasing).
- You have $\{\text{MAX_LOOP}-1\}$ questioning rounds available: the early rounds focus on global perception, the middle rounds gather key evidence, and the final rounds perform verification questioning.

[Output Format (Strict Requirements)]

- If further questioning is needed:

Start with `<thinking>...</thinking>` (briefly explain the reasoning and purpose of the next question), then output **only one** `<question>...</question>`.

- If ready to give the final answer:

Start with `<thinking>...</thinking>` (summarize key evidence and note that final checks have been completed), then output `\{\text{FINAL_PREFIX_EN}\} \dots`

- Each round must **begin** with `<thinking>...</thinking>` and be followed by **exactly one** of the two options: `<question>...</question>` or `\{\text{FINAL_PREFIX_EN}\} \dots`. No other content is allowed.

- Inside `<thinking>`, do **not** mention external entities such as “the perception model,” “user,” or “conversation.”

- Use **English** for internal reasoning and questioning, but ensure that the **final answer matches the input query’s language**.

Figure 1. The system prompt for the Reasoner in SocraticAgent.

4. RL Reward Function

4.1. Grounding Reward

For the grounding task, the model is required to output a bounding box in the form $[x1, y1, x2, y2]$ after a complete `<think></think>` block. Our reward contains two components: an IoU-based accuracy term and a lightweight format term.

Format reward. For the grounding task, we apply a lightweight *format reward* to encourage the model to produce both a well-formed reasoning block and a structured bounding-box output. We evaluate two binary indicators: (i) whether the response contains a complete, matched `<think></think>` block, and (ii) whether a valid four-number bounding-box list of the form $[x1, y1, x2, y2]$ appears after the closing `</think>` tag. We denote these indicators as s_{think} and s_{bbox} , respectively. The format

System Prompt for the Perceiver

You are an image interpretation expert collaborating with a **reasoning model that has very weak logical ability**. Together, through multi-turn dialogue, you will complete **general remote sensing tasks** (classification/attribute, localization/counting, relation/change analysis, VQA, etc.). The reasoning model can **only understand the textual descriptions** of your perception results — it **cannot see the image directly**. Therefore, you must respond to each of its questions about the image **accurately and completely**, without adding any information that is irrelevant to the question. Your tone should resemble a **natural inner monologue** of a person carefully observing an image. Always begin your response with: "Let's look at the image," and then continue with your detailed observation.

Figure 2. The system prompt for the Perceiver in SocraticAgent.

System Prompt for the Verifier

You are a **strict answer evaluator**. Given a Query, Answer, and GT, output only:
1) "ACCEPT"
2) "REJECT: <brief reason>"

Figure 3. The system prompt for the Verifier in SocraticAgent.

reward is then defined as

$$r_{\text{fmt}} = \frac{s_{\text{think}} + s_{\text{bbox}}}{2},$$

IoU accuracy. We first normalize the response and verify that a well-formed reasoning block (i.e., a matched pair `<think>...</think>`) exists. Only the text appearing *after* the closing tag `</think>` is used for prediction parsing. From this tail segment, we extract the first valid four-number list `[x1, y1, x2, y2]`; if no valid list is found, the prediction is treated as invalid. The accuracy score is then computed using the IoU between the predicted and ground-truth boxes directly in the original pixel coordinate system:

$$r_{\text{acc}} = \begin{cases} \text{IoU}(\text{bbox}_{\text{pred}}, \text{bbox}_{\text{gt}}), & \text{if a valid box is parsed,} \\ 0, & \text{otherwise.} \end{cases}$$

Overall reward. The final grounding reward combines IoU accuracy and the format term:

$$r_{\text{overall}} = (1 - \lambda)r_{\text{acc}} + \lambda r_{\text{fmt}},$$

where we use $\lambda = 0.1$ in all experiments. Each sample is scored independently, and we log r_{overall} , IoU, r_{acc} , and r_{fmt} during training.

4.2. Multiple-Choice VQA Reward

For the multiple-choice VQA task, the model predicts a subset of option letters (e.g., A, C, D) after completing the `<think></think>` reasoning block. Only the text following the last `</think>` tag is considered for answer extraction. We first identify the last line in this tail that contains independent letter tokens; if none are found, we treat the prediction as invalid.

Given the ground-truth correct set $G \subseteq \mathcal{A}$ and the model's predicted set $P \subseteq \mathcal{A}$, where \mathcal{A} is the set of all allowed options, we compute a symmetric, option-level accuracy score. Each option contributes equally: selecting a correct option and correctly rejecting an incorrect option both receive positive credit, while selecting an incorrect option or failing to select a correct one is penalized. The accuracy reward is therefore:

$$r_{\text{acc}} = 1 - \frac{1}{|\mathcal{A}|} \sum_{a \in \mathcal{A}} |\mathbf{1}\{a \in G\} - \mathbf{1}\{a \in P\}|.$$

If the model outputs any letter not contained in \mathcal{A} , we apply a hard-zero accuracy (i.e., $r_{\text{acc}} = 0$), matching our implementation.

Format reward. The VQA setting uses the same lightweight format reward as the grounding task. Because empty outputs are allowed and interpreted as selecting no options, the format reward does not require the predicted letters to appear after the `</think>` tag. Instead, we simply check whether the model produces a complete and matched `<think></think>` reasoning block. If the block is present, the model earns a format score of 1; otherwise, the score is 0.

Overall reward. The final VQA reward is a convex combination of accuracy and format:

$$r_{\text{overall}} = (1 - \lambda)r_{\text{acc}} + \lambda r_{\text{fmt}},$$

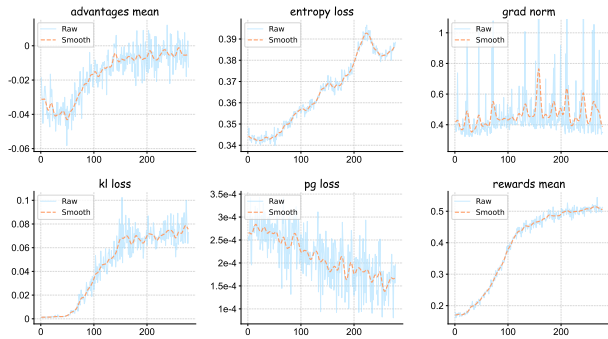
with $\lambda = 0.1$ throughout all experiments.

This reward design provides a graded, stable learning signal across all options without relying on task-specific heuristics, while the matched `<think></think>` constraint ensures structured reasoning outputs.

5. RL Training Dynamics Curves

Figure 4 visualizes the evolution of key optimization statistics during the two RL stages in our pipeline: RL-Grounding and RL-VQA. The top block corresponds to the RL Grounding stage and the bottom block to the RL-VQA

RL-Grounding Training Dynamics



RL-VQA Training Dynamics

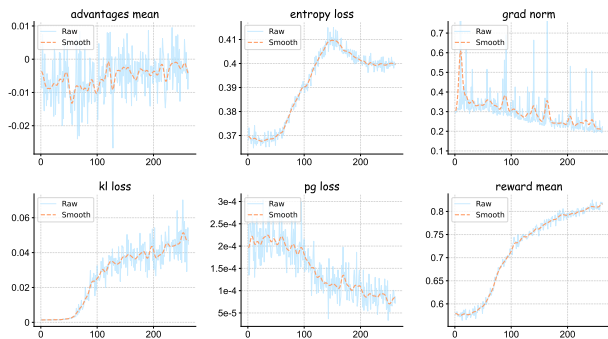


Figure 4. RL training dynamics for our two-stage pipeline. Top: RL-Grounding stage; bottom: RL-VQA stage. Each panel reports the evolution of a different statistic (advantage mean, entropy loss, gradient norm, KL loss, policy-gradient loss, and reward mean), with light curves showing raw values and dashed curves showing Gaussian-smoothed trends over training steps.

stage; in both cases we plot the same set of metrics, including mean advantage, actor gradient norm, entropy loss, KL loss, policy gradient loss, and average reward. For each quantity, we plot both the raw measurements and their Gaussian-smoothed trends, providing a compact view of how the GRPO optimization behaves over the training steps in both stages.

6. Difference Between Multiple-Choice VQA and Standard VQA

To assess the effectiveness of our proposed multiple-choice reformulation of VQA, we additionally perform an ablation study using the original dataset and model settings, but applying reinforcement learning directly on the *standard* free-form VQA answers. This experiment allows us to isolate and compare the impact of our multiple-choice VQA design against the conventional VQA supervision.

We first evaluate the final model performance on the

Table 1. Ablation study on VQA data reconstruction strategies on the RSVQA. We compare our proposed Multiple-Choice VQA reconstruction against the Standard VQA (Free-form) supervision. Our method achieves consistently superior performance across all metrics.

| Method | Avg@5 | Conv@5 | Pass@5 |
|-------------------------------|--------------|--------------|--------------|
| Standard VQA (Free-form) | 74.73 | 76.86 | 90.91 |
| Ours (Multiple-Choice) | 75.16 | 78.29 | 92.51 |

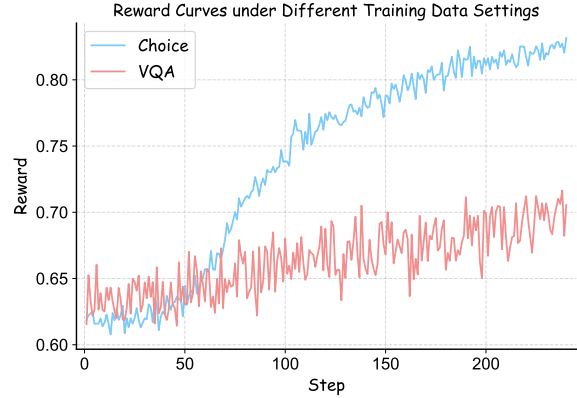


Figure 5. Ablation comparing reinforcement learning on the original free-form VQA answers (red) versus our reconstructed multiple-choice VQA formulation (blue). Standard VQA produces unstable and oscillatory training dynamics, while the multiple-choice approach yields smooth and efficient reward improvement.

RSVQA. As presented in Table 1, our multiple-choice reconstruction strategy achieves consistently superior results across all metrics compared to the standard free-form supervision. To understand the source of this performance gap, we visualize the training stability in Figure 5. When trained with standard free-form VQA answers, the model exhibits slow reward improvement and frequent large-magnitude oscillations during GRPO optimization. This indicates that the original VQA supervision produces an unstable and inefficient reward signal for RL, making it difficult for the model to learn consistent behaviors. In contrast, our multiple-choice VQA reformulation leads to a much smoother and steadily increasing reward curve under the *same* RL setup. This demonstrates that transforming VQA into a structured multi-option prediction task greatly stabilizes the reward landscape and improves training efficiency.

Overall, the ablation confirms that the multiple-choice VQA formulation provides a more reliable and effective RL objective compared to standard VQA, and is essential for achieving stable reinforcement learning in our setting.

7. Additional Experimental Results

In this section, we collect the additional experimental results that were used to further analyze the RS-EoT paradigm, including comparisons with remote-sensing VLMs, classification and cross-modality transfer results, an ablation on the order of the two RL stages, Pass@1 evaluation, and a brief discussion of common failure modes.

7.1. Comparison with Remote-Sensing VLMs

RS-EoT-7B is a reasoning-oriented model and is therefore most directly compared with other remote-sensing multi-modal models on the FIT-RSFG-VQA benchmark. Table 2 reports additional results for GeoChat and EarthDial. Compared with these instruction-style baselines, RS-EoT-7B achieves substantially better performance on all reported metrics, indicating that the explicit reasoning traces learned by RS-EoT translate into stronger downstream reasoning ability on fine-grained remote-sensing VQA.

Table 2. Additional comparison with remote-sensing VLMs on FIT-RSFG-VQA.

| Model | Avg@5 | Conv@5 | Pass@5 |
|-----------|--------------|--------------|--------------|
| GeoChat | 53.89 | 52.39 | 54.66 |
| EarthDial | 53.51 | 53.87 | 75.50 |
| RS-EoT-7B | 67.85 | 68.90 | 85.28 |

7.2. Classification and Cross-Modality Results

Besides general VQA and grounding, we also evaluate RS-EoT-7B on classification benchmarks and on a SAR benchmark to examine whether the learned reasoning strategy transfers beyond the main training tasks. As shown in Table 3, RS-EoT-7B consistently outperforms the base Qwen2.5-VL-7B model on RSFG-SC, UC_Merced, and OpenSARUrban. These results suggest that the reasoning-oriented post-training procedure does not only help on VQA-style tasks, but also improves classification performance and exhibits reasonable transfer to non-RGB remote-sensing imagery.

Table 3. Evaluation on classification benchmarks and a SAR benchmark.

| Model | RSFG-SC | UC_Merced | OpenSARUrban (SAR) |
|---------------|--------------|--------------|--------------------|
| Qwen2.5-VL-7B | 57.42 | 69.57 | 25.78 |
| RS-EoT-7B | 64.05 | 72.62 | 26.34 |

7.3. Ablation on the Order of the Two RL Stages

We further study the order of the two RL stages in our training pipeline. Table 4 compares three variants: performing RL-Grounding before RL-VQA (VG+VQA), reversing the order (VQA+VG), and merging both types of data into a single RL stage (Merged). The VG+VQA strategy yields the best overall trade-off, especially on grounding metrics,

which indicates that grounding-first RL provides a stronger perceptual foundation for the later reasoning-heavy VQA stage. This observation is consistent with the motivation of our pipeline design: cold-start SFT is used to initialize the reasoning format, whereas RL on grounding data supplies the fine-grained visual alignment needed for subsequent complex reasoning.

Table 4. Ablation study on the order of the two RL training stages.

| Model | RSFG-VQA | | | VRSBench-ref | | |
|--------|--------------|--------------|--------------|--------------|--------------|--------------|
| | Avg@5 | Conv@5 | Pass@5 | IoU@50 | IoU@70 | mIoU |
| VG+VQA | 67.85 | 68.90 | 85.28 | 54.71 | 32.40 | 48.04 |
| VQA+VG | 65.57 | 68.75 | 83.64 | 51.42 | 29.10 | 45.77 |
| Merged | 65.88 | 69.28 | 83.55 | 51.06 | 29.02 | 45.52 |

7.4. Pass@1 Evaluation

For completeness, Table 5 reports Pass@1 results on three benchmarks. RS-EoT-7B improves over the Qwen2.5-VL-7B baseline in all cases, showing that the gain of our method is not limited to best-of- k sampling and is already visible under single-sample decoding.

Table 5. Pass@1 evaluation results.

| Model | RSFG-VQA | VRSBench | RSVQA |
|---------------|--------------|--------------|--------------|
| Qwen2.5-VL-7B | 62.42 | 61.97 | 67.39 |
| RS-EoT-7B | 68.16 | 63.33 | 75.08 |

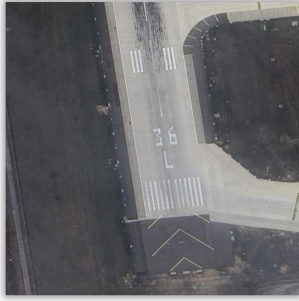
7.5. Observed Failure Modes

We observe that RS-EoT still struggles in scenes containing many dense, small-scale objects and in queries that require exact counting under heavy occlusion or clutter. In such cases, the reasoning chain may remain well-structured while the underlying visual evidence is incomplete or imprecise. These examples indicate that better fine-grained perception and stronger object enumeration remain important directions for improving reasoning models in remote sensing.

8. Case Study

We provide additional qualitative examples to further demonstrate the effectiveness of RS-EoT-7B in complex remote sensing reasoning scenarios. Specifically, we present extended case studies covering both Remote Sensing General VQA tasks (Fig. 6, Fig. 7, and Fig. 8) and Fine-grained Grounding tasks (Fig. 9 and Fig. 10). These visualizations illustrate how the model employs the iterative evidence-seeking paradigm to achieve robust reasoning and precise localization across diverse queries.

Q: What is the runway designation shown in the image?



Alright, I will reason in a self Q&A style and give the final reply. Start with a global check to confirm the scene contains runway-like infrastructure before focusing on specific markings. Is there a long, wide, straight paved strip visible in the image? Let's look at the image. Yes, there is a prominent long, wide, straight paved strip visible in the image. It is a gray surface with white and yellow markings, running diagonally across the image from the top-center towards the bottom-left. Focus on the strip's markings to identify the runway designation associated with the visible numbers. Are there large white runway numbers visible on the paved strip? Let's look at the image, yes, there are large white runway numbers visible on the paved strip. Specifically, the number "36" is clearly visible, with an additional "L" next to it. Global view confirmed a long, wide runway; detailed check verified large white runway numbers "36" with an additional "L." No inconsistencies detected; this is a standard runway designation. Final review complete.</think> 36L

Q: Are there any vehicles about to enter the ramp?



Alright, I will reason in a self Q&A style and give the final reply. Start with a global check to confirm the presence of a ramp-like road connection, which is essential before assessing vehicles entering it. Is there a distinct road ramp connection visible, such as an on-ramp or off-ramp, merging into the main highway? Let's look at the image. Yes, there is a distinct road ramp connection visible. A curved road, located in the bottom right portion of the image, merges into the main highway. This ramp is clearly separated from the main highway by a grassy area, and it features white road markings, including chevron-like shapes at its entrance, indicating it's a ramp connection. With the ramp connection confirmed, focus on whether any vehicles are positioned on it, which directly addresses the query about vehicles entering the ramp. Are there any vehicles visible on the ramp road near the ramp entrance? Let's look at the image. Yes, there is one vehicle visible on the ramp road near the ramp entrance. It is a white vehicle, positioned on the lane that curves to the right in the bottom right portion of the image. The scene includes a distinct ramp connection, and one white vehicle is visible on the ramp near the entrance, which qualifies as a vehicle about to enter the ramp. No contradictions found upon review.</think> Yes

Q: If I start from the yellow construction site in the lower-left corner and walk along the road to the parking lot in the upper-right corner, how many intersections will I pass through?



Alright, I will reason in a self Q&A style and give the final reply. Start with a global check to confirm the presence and location of the target yellow construction site to anchor the walk route. Is there a yellow construction site located in the lower-left corner of the image? Let's look at the image. Yes, there is an area that appears to be a yellow construction site located in the lower-left corner of the image. It's a large, irregularly shaped area with a light brown, earthy tone, distinct from the surrounding grey roads and building rooftops. We have confirmed the starting point (yellow construction site) is in the lower-left and the destination (parking lot) is in the upper-right. Next, to mentally map the walk route, we should verify whether there is a major road running vertically near the center, which would align with a likely walking path toward the upper-right corner. Is there a prominent vertical road running near the center of the image? Let's look at the image, yes, there is a prominent vertical road running near the center of the image. It's a wide, dark gray paved road that extends from the top edge of the image all the way down to the bottom edge, passing through the

middle section. We have identified the starting point (yellow construction site, lower-left) and the general destination (parking lot, upper-right). A prominent vertical road runs through the center, so the most direct path toward the upper-right will cross this central road. To determine the exact count, we need to know how many times the central road is crossed along the way. How many times does the prominent vertical road intersect other roads along the route from the yellow construction site to the parking lot in the upper-right corner? Let's look at the image. I need to identify the prominent vertical road, locate the yellow construction site, and a parking lot in the upper-right corner. Then I will count how many times the prominent vertical road intersects other roads along that route.1. ****Locating the Prominent Vertical Road****: The most prominent vertical road is a wide, dark gray road that runs from the top of the image down towards the bottom, passing through the center.2. ****Identifying the Yellow Construction Site****: The yellow construction site is located in the bottom-left corner of the image, characterized by a large, irregular area of bare earth with some yellow markings.3. ****Identifying a Parking Lot in the Upper-Right Corner****: There is a large parking lot with many cars in the upper-right portion of the image, located near several buildings with light-colored roofs. Now, I will mentally trace the route from the yellow construction site to the upper-right parking lot along the prominent vertical road and count the intersections. Starting from the bottom-left corner (yellow construction site), the route goes straight up along the prominent vertical road. Along this path, the first road encountered and crossed above the main vertical road is a horizontal road that runs roughly from the left edge to the right edge, intersecting the prominent vertical road in the middle-left area of the image. This is the first intersection. After crossing this first horizontal road, the route continues straight up along the prominent vertical road. The next road encountered and crossed above this point is another horizontal road that intersects the prominent vertical road in the middle-right area. This is the second intersection. Finally, the route reaches the upper-right area of the image, where there is a large parking lot (the destination). This parking lot is located on a corner, bordered by roads on its right and top sides. The prominent vertical road runs along the left and bottom edges of this parking lot area. Although it is a large area, it does not appear to be a distinct "parking lot" that crosses another major road; rather, it is a large lot situated between two intersecting roads. Therefore, no additional intersections with other roads are gained by reaching this parking lot. Summarizing the intersections: * First intersection: horizontal road at middle-left. * Second intersection: horizontal road at middle-right. * Arrive upper-right parking lot area, but no additional road crossings beyond these two.</think> 2

Figure 6. Reasoning cases of RS-EoT-7B (Part 1).

Q: How many cars are in the left-turn lane?



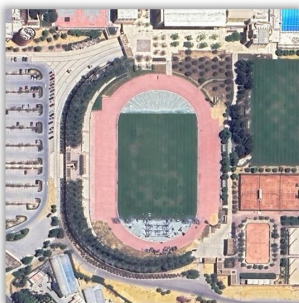
Alright, I will reason in a self Q&A style and give the final reply. Start with a global observation to confirm the presence of a left-turn lane at the intersection, establishing the target lane's identity before counting. Is there a left-turn lane visible at the intersection? Let's look at the image. Yes, there is a left-turn lane visible at the intersection. It is the lane on the far left side of the road, marked with a white arrow pointing towards the upper left corner of the image. We have identified the left-turn lane. Next, determine whether any vehicles are present in that specific lane to set the base count before checking for additional cars. Is there at least one car positioned within the left-turn lane at the intersection? Let's look at the image. Yes, there is at least one car positioned within the left-turn lane at the intersection. I can see a dark-colored car in the leftmost lane, which is the left-turn lane, near the top of the image. Global observation confirmed a left-turn lane at the intersection. Targeted check indicated at least one car present within that lane. No conflicting cues or indications of an empty lane were observed. Final review finds no additional cars requiring counting.</think> 1

Q: What is the main land-cover type in this remote sensing image?



Alright, I will reason in a self Q&A style and give the final reply. Start with a global assessment to identify the single most dominant land-cover category by area. Determine if water bodies cover the largest area to guide subsequent focused checks. Does a large water body (river, lake, or sea) cover the largest area in the image? Let's look at the image. The image is primarily covered by dense green vegetation, which appears to be trees and bushes. There are also numerous small, irregular patches of lighter brown or greyish ground visible throughout the vegetation, which seem to be clearings or unpaved areas. Additionally, there are many residential buildings and paved roads. There are some very small, dark blue or greenish features, which could be swimming pools, located next to some of the buildings in the lower right and central-right portions of the image. However, these are very small in area compared to the vast expanse of vegetation. Therefore, no large water body such as a river, lake, or sea covers the largest area in the image. The largest area is covered by vegetation. Global observation indicates that vegetation covers the largest area, with no large water bodies and only small pools present. Final review finds no other land-cover type dominating.</think> Vegetation

Q: Describe this remote sensing image.



Alright, I will reason in a self Q&A style and give the final reply. Start with a broad global assessment to determine the overall scene type and guide subsequent focused questions. Is the scene predominantly urban/built-up? Let's look at the image. The scene contains a large stadium with a running track and a green field, multiple tennis courts, a basketball court, a parking lot with numerous cars, and many buildings with roads and paved areas. While there are green spaces and trees, a significant portion of the image is covered by man-made structures and infrastructure. Therefore, the scene is predominantly urban/built-up. From the global view, the scene appears urban with a prominent sports complex. Next, confirm the central anchor object to structure the caption. Is there a large stadium with a green field and a running track at the center of the image? Let's look at the image, yes, there is a large stadium with a green field and a running track at the center of the image. We have a global understanding: an urban scene centered on a stadium with sports facilities and parking. Next, confirm surrounding sports facilities for detailed spatial description. Are multiple tennis courts visible adjacent to the stadium? Let's look at the image. Yes, there are multiple tennis

courts visible adjacent to the stadium. They are located to the right of the stadium in the image. We have a global understanding and confirmed the stadium and adjacent sports facilities. Next, determine the parking arrangement to describe spatial flow and context. Is there a large parking lot with many cars immediately to the left of the stadium? Let's look at the image. Yes, there is a large parking lot with many cars immediately to the left of the stadium. We have identified an urban scene centered on a stadium with adjacent sports facilities and parking. To enrich the caption with architectural context, determine the predominant building type around the stadium. Are the surrounding buildings predominantly low-rise with flat roofs? Let's look at the image. Yes, the surrounding buildings appear to be predominantly low-rise with flat roofs. There are no visible indications of significant height or sloped/pitched roofs on the buildings. Global view shows a predominantly urban scene. Confirmed a central stadium with a green field and running track, adjacent sports facilities (tennis courts to the right), a large parking lot with many cars to the left, and a ring of low-rise buildings with flat roofs. Roads and paved areas are extensive; vegetation is secondary. No contradictions found; ready to deliver the caption.</think> A high-resolution urban scene centered on a multi-purpose stadium complex. At the heart of the image is a rectangular green playing field encircled by a red running track, with tiered spectator stands surrounding it. Immediately to the right of the stadium lie several tennis courts and a basketball court, set within paved aprons and bordered by paved roads. To the left of the stadium, a large parking lot packed with many cars is connected by a network of access roads. The area is densely built, with numerous low-rise buildings featuring flat roofs forming a continuous urban fabric around the sports facilities. Surfaces vary from dark asphalt and light concrete to red track surfaces and bright green turf, with trees and planted greenery lining pavements and interspersed between buildings, but vegetation is secondary to built-up areas and paved surfaces. The overall layout exhibits a clear contrast between the stadium's oval athletics track and the rectilinear arrangement of courts, parking, and buildings, all organized by a visible road grid.

Figure 7. Reasoning cases of RS-EoT-7B (Part 2).

Q: How many vessels are sailing in the image?



Alright, I will reason in a self Q&A style and give the final reply. Start with a global check to confirm the presence of a water body, which guides where vessels would appear. Is there a large water body visible in the image? Let's look at the image. Yes, there is a large water body visible in the image, occupying a significant portion of the lower and central areas. Shift focus to counting by spatial partition to efficiently determine the total number of vessels. How many vessels are visible in the left half of the image? Let's look at the image. I can see several vessels on the water. I will mentally divide the image vertically into two halves. In the left half of the image, I observe three vessels. These are long, narrow, dark-colored boats, each with a visible wake trailing behind them. In the right half of the image, I observe four vessels. These vessels are also long, narrow, and dark-colored, with visible wakes. Therefore, there are 3 vessels in the left half of the image. We have a global understanding and a partial count (3 in the left half). To compute the total, we need the count in the right half. How many vessels are visible in the right half of the image? Let's look at the image. I can see several vessels on the water. I'll divide the image vertically into two equal halves to count the

vessels in the right half. In the right half of the image, I can identify five vessels: 1. A long, dark vessel with a visible wake, located towards the bottom right. 2. Another long, dark vessel with a visible wake, positioned slightly above the first one and to the left. 3. A shorter, dark vessel with a visible wake, located further up and to the left from the second one. 4. A very small, dark speck in the water, located slightly above the third vessel and to the left. 5. Another very small, dark speck in the water, located further up and to the left from the fourth vessel. Therefore, there are 5 vessels visible in the right half of the image. Global scan confirms a large water area. Focused counts: 3 vessels in the left half and 5 in the right half, totaling 8. No contradictions detected; final review suggests no vessels near the vertical midline that would cause double counting. 8

Q: Provide a detailed caption.

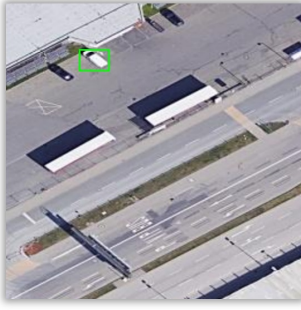


Alright, I will reason in a self Q&A style and give the final reply. Begin with global scene categorization to establish whether the landscape is predominantly urban, which guides subsequent focused questions on infrastructure, water, and land-cover patterns. Is the scene predominantly urban with dense building coverage? Let's look at the image. The scene is not predominantly urban with dense building coverage. Instead, it features large, light-colored sandy areas, several large triangular structures that appear to be pyramids, and a network of roads. There are also some scattered, smaller, darker structures, possibly related to the pyramids or other facilities, but these are not dense and do not constitute a predominantly urban environment. From the initial global view, the scene includes light sandy terrain, several large pyramids, and a road network with sparse darker structures. To craft a precise caption, first confirm the exact number of pyramids as a key scene element. How many large pyramidal structures are visible in the scene? Let's look at the image, I can see three large pyramidal structures. One is in the upper-middle section, another is in the lower-middle section, and a third is partially visible in the bottom-left corner. To enrich the caption with

environmental context, determine the dominant surrounding land cover beyond the pyramids and roads. Is the ground surface predominantly light-colored sand with broad sandy expanses? Let's look at the image, yes, the ground surface is predominantly light-colored sand, with broad expanses of it covering most of the visible area. To enrich the caption with environmental context, confirm whether any water features are present, which would significantly affect scene description. Is any water body visible in the scene? Let's look at the image, I have carefully examined the entire scene, and I do not see any visible water bodies. To enrich the caption with contextual infrastructure, verify whether there are additional man-made structures beyond the pyramids and roads, which helps describe the site's layout and land use. Besides the pyramids and roads, are there any other notable man-made structures visible? Let's look at the image. Yes, besides the pyramids and roads, there are other notable man-made structures visible. In the upper right portion of the image, there is an area with numerous small, dark, rectangular shapes arranged in rows. These appear to be buildings or paved areas. Some of these are larger, suggesting they could be facilities or compounds. This area is adjacent to what looks like an unpaved or dirt track. Further down, towards the middle right, there's a larger, irregularly shaped complex of darker material, possibly debris or construction refuse, located near some sparse vegetation. In the bottom right corner, there's a small, bright, rectangular object that could be a vehicle or a small piece of equipment. Global view shows a non-urban desert scene dominated by light sand, three large pyramids, and a road network; detailed checks confirm three pyramids (one upper-middle, one lower-middle, one partial bottom-left), broad sandy expanses, no water bodies, additional man-made clusters of small dark rectangular structures (upper right), an irregular darker refuse/ debris area (middle right), sparse vegetation, and a bright rectangular object (bottom right). All key regions reviewed and consistent; ready to provide the caption. A remote desert landscape dominated by light tan sand dunes and broad, open terrain, punctuated by three large pyramidal structures: one in the upper-middle, one in the lower-middle, and a partial third at the bottom-left edge. A network of roads and unpaved tracks weaves around the pyramids, linking them to clusters of small, dark rectangular buildings and paved areas in the upper-right portion of the scene. The area also contains an irregular, darker accumulation of refuse or debris near some sparse vegetation on the right, with a few isolated vehicles or small pieces of equipment visible, including a bright rectangular object in the bottom-right corner. No water bodies are present, and the surrounding ground is largely unvegetated, with strong sunlight textures and sharp shadows highlighting the pyramids and structural clusters.

Figure 8. Reasoning cases of RS-EoT-7B (Part 3).

Q: Locate the white car in the image. Return the coordinates as "[x1, y1, x2, y2]".



Alright, I will reason in a self Q&A style and give the final reply. Start with a global check to understand the scene layout and identify likely regions where a car might appear, such as roads or parking areas. Is there a paved road or parking area visible in the image? Let's look at the image. Yes, there are several paved areas visible. There's a large paved area in the upper left with a white car parked on it, and multiple paved roads with lane markings running diagonally across the lower and middle parts of the image. We have identified a large paved area in the upper-left where a white car is likely parked. Next, confirm the car's presence in that area. Is there a white car parked on the large paved area in the upper-left quadrant? Let's look at the image. Yes, there is a white car parked on the large paved area in the upper-left quadrant of the image. Global observation revealed paved roads and a large paved area; focused observation confirmed a white car parked on the upper-left paved area. No conflicting cues found upon final review; coordinates will be returned in a self Q&A style.</think> [128,81,177,115]

Q: Locate the swimming pool in the image. Return the coordinates as "[x1, y1, x2, y2]".



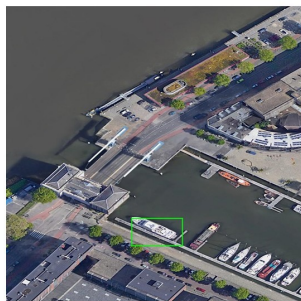
Alright, I will reason in a self Q&A style and give the final reply. Start with a global scan to detect any prominent water-like area, which is a common high-level cue for pools. This broad check guides subsequent focused localization. Is there any noticeable blue or turquoise area with smooth, curved boundaries visible in the image? Let's look at the image, yes, there is a very noticeable blue area with smooth, curved boundaries visible in the image. It is a swimming pool located in the upper-middle portion of the image. We have a global indication of a blue, curved-boundary area likely being the pool. To confirm precise localization for coordinates, first verify that there is only one such area to avoid ambiguity. Is there exactly one blue area with smooth, curved boundaries in the image? Let's look at the image, I can see a prominent blue area with smooth, curved boundaries. This area is a swimming pool. I do not observe any other blue areas with smooth, curved boundaries in the image. Global observation identified a single blue, smooth-boundary area; focused checks confirmed it is a swimming pool with curved edges; no conflicting cues found; final review complete.</think> [369,203,538,297]

Q: Locate the train in the image. Return the coordinates as "[x1, y1, x2, y2]".



Alright, I will reason in a self Q&A style and give the final reply. Start with a global check to see if any train-like object exists in the scene before narrowing down its location. Is there at least one train visible anywhere in the image? Let's look at the image. Yes, there is at least one train visible in the image. It is located on the far right side, towards the top, on the railway tracks. A portion of a train, likely its roof, can be seen extending into the image frame from the right edge. Global observation confirmed the presence of a train; it was noted near the top-right along the tracks. No conflicting cues found upon quick review. Ready to provide the concise location.</think> [753,127,796,309]

Q: Locate the largest cruise ship in the image. Return the coordinates as "[x1, y1, x2, y2]".



Alright, I will reason in a self Q&A style and give the final reply. Start with a global scan to confirm the presence of maritime context, then narrow to locating large ship-like vessel(s) in the main water area. Is there a large water body visible in the image? Let's look at the image. Yes, there is a large water body visible in the image, occupying a significant portion of the left and central areas. From the global scan, a large water body is present. Next, I need to localize large ship-like vessels to identify potential candidates for the largest cruise ship. Are there any large ship-like vessels visible on the water? Let's look at the image. Yes, there are several large ship-like vessels visible on the water. Specifically, in the lower central part of the image, there is a very prominent, long, white vessel docked alongside a pier. It has a distinct hull and a superstructure, making it appear like a large yacht or a private vessel. Further down and to the right on the water, there's a cluster of smaller, more traditional boat-like shapes, which appear to be barges or working vessels, but one of them is quite long and somewhat rectangular, resembling a barge. In the upper left portion of the water, there's a long, narrow structure, which appears to be a bridge or a drawbridge, extending over the water. It's a large, man-made structure, but it's a bridge rather than a ship. To identify the largest cruise ship, I will first confirm whether the prominent long white vessel in the lower center is a passenger-carrying cruise ship rather than a barge. Does the long white vessel in the lower center have multiple passenger cabins visible on its deck? Let's look at the image. The long white vessel in the lower center of the image is a large yacht. On its deck, multiple passenger cabins are visible. Global scan confirms a large water body with several ship-like vessels. Focused inspection identifies a prominent long white vessel with multiple cabins in the lower center, consistent with a cruise ship. No larger candidate observed; final checks complete.</think> [341,575,476,647]

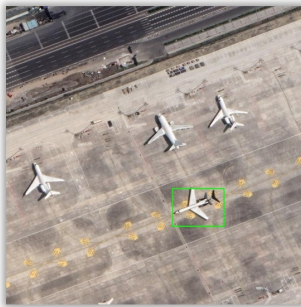
Figure 9. Reasoning cases of RS-EoT-7B (Part 4).

Q: Locate **the truck** in the image. Return the coordinates as "[x1, y1, x2, y2]".



Alright, I will reason in a self Q&A style and give the final reply. Begin with a global check to confirm the presence of a truck before localizing it. Is there at least one truck visible in the image? Let's look at the image. Yes, there is at least one truck visible in the image. It is located on the road in the bottom left portion of the image, appearing as a large white and red vehicle. Presence of a truck is confirmed. To provide precise localization for the coordinates, verify whether there is exactly one truck in the image so that global and local positions align. Is there exactly one truck visible in the image? Let's look at the image. Yes, there is exactly one truck visible in the image. It is located on the road in the bottom left portion of the image. Global observation confirmed presence of a truck; focused check verified there is exactly one truck visible. Final review finds no additional trucks elsewhere. Ready to provide the coordinates.</think> [56,594,154,650]

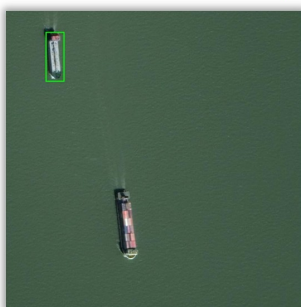
Q: Locate **the taxiing plane** in the image. Return the coordinates as "[x1, y1, x2, y2]".



Alright, I will reason in a self Q&A style and give the final reply. Start with a global check to confirm the presence of airplanes before focusing on taxiway localization. Are there any airplanes visible in the image? Let's look at the image. Yes, there are airplanes visible in the image. I can see four distinct airplanes. We have confirmed multiple airplanes are present. Next, determine whether one is on a taxiway rather than at a terminal gate, which is necessary to identify the "taxiing plane." Is there an airplane positioned on a taxiway (not at a terminal gate)? Let's look at the image. Yes, there is an airplane positioned on a taxiway. The large paved area where the airplanes are located is a taxiway, and there are no visible terminal gates in the image. We identified multiple airplanes and confirmed one is on a taxiway. To provide precise coordinates, we should locate the taxiway-facing nose of that specific plane. Is the airplane on the taxiway oriented with its nose pointing toward the top edge of the image? Let's look at the image. There are four airplanes visible. The airplane on the far left is oriented with its nose pointing towards the top-left corner of the image. The second airplane

from the left is oriented with its nose pointing towards the right-middle of the image. The third airplane from the left is oriented with its nose pointing towards the bottom-right of the image. The airplane on the far right is oriented with its nose pointing towards the top-right corner of the image. Therefore, only two of the airplanes are oriented with their noses pointing toward the top edge of the image. We have multiple airplanes on a taxiway, with at least two oriented toward the top edge. To uniquely identify the "taxiing plane," I will confirm which one is actively moving by checking for a trailing wake. Is there a plane showing a visible wake trailing behind it, indicating movement? Let's look at the image. I am examining all four airplanes visible in the image. I do not see any visible wake trailing behind any of the planes. All the airplanes appear to be stationary on the tarmac. There seems to be a mismatch: earlier indications suggested at least one plane is on a taxiway, but the orientation check showed most noses aren't toward the top edge, and no wake was seen. To resolve this, I will verify whether any plane's nose points toward the bottom edge, which would indicate taxiing. Is any airplane oriented with its nose pointing toward the bottom edge of the image? Let's look at the image. Yes, there is one airplane oriented with its nose pointing toward the bottom edge of the image. This airplane is located in the bottom right portion of the image. Global view confirmed multiple airplanes on a taxiway. Detailed checks showed most noses aren't toward the top edge, but one plane in the bottom-right has its nose pointing toward the bottom edge, consistent with taxiing. Final review finds no conflicting cues.</think> [447,490,583,589]

Q: Locate **the smaller ship** in the remote sensing image. Return the coordinates as "[x1, y1, x2, y2]".



Alright, I will reason in a self Q&A style and give the final reply. Start with a global assessment to understand the scene layout and identify whether water dominates, which guides where ships would appear. Is the majority of the image covered by water? Let's look at the image. The majority of the image is covered by a dark green, textured surface that appears to be water. Yes, the majority of the image is covered by water. With the global scene confirmed as mostly water, next I need to verify that ships are present to locate the smaller one. Are there multiple ships visible on the water surface? Let's look at the image. Yes, there are multiple ships visible on the water surface. I can clearly see two distinct ships. Global scene is water with two ships. To identify which is smaller, I will compare their sizes directly. Is the ship in the upper-left part of the image smaller than the ship in the lower-central part? Let's look at the image. The ship in the upper-left part of the image is smaller than the ship in the lower-central part. Global view shows a mostly water scene with two ships. Focused comparison confirms the upper-left ship is smaller than the lower-central one. No contradictions observed; final check complete.</think> [108,58,154,187]

Figure 10. Reasoning cases of RS-EoT-7B (Part 5).