SegEarth-R2: Towards Comprehensive Language-guided Segmentation for Remote Sensing Images

Supplementary Material

7. More information of LaSeRS

7.1. Definition of Four Dimensions of LaSeRS

Hierarchical Granularity. A primary challenge in RS arises from the vast and complex nature of the imagery. A single scene often contains a diverse array of objects at vastly different scales, from large-scale collections (*e.g.*, "all indistinguishable vehicles densely parked in the parking lot") to specific instances (*e.g.*, "a cargo truck parked in the bottom left"). Consequently, language queries used to describe these scenes inherently exhibit hierarchical granularity. As shown in Figure 1, this hierarchical granularity can be categorized into conceptual and segmentation granularity. The first, hierarchical concept granularity, relates to the level of semantic abstraction in the query, requiring the model to understand a hierarchy from coarse categories (*e.g.*, "airplane") to fine-grained sub-classes (*e.g.*, "Boeing 787"). This is particularly challenging in RS images due to noise, small objects, and the need for expert knowledge. The second, hierarchical segmentation granularity, dictates the spatial scope of the desired output mask, which operates at three distinct levels: semantic (*e.g.*, "all tennis courts"), instance (*e.g.*, "the rightmost tennis court"), and part (*e.g.*, "the service area of the rightmost tennis court").

Target Multiplicity. This challenge arises from queries that reference multiple objects simultaneously, testing a model's ability to parse and ground complex instructions. This is illustrated by the target multiplicity and long query cases in Figure 1.

Reasoning Requirements. This challenge spans a spectrum from explicit to implicit reasoning [28]. Explicit queries relate to literal visual attributes (*e.g.*, "a large ground track field"). In contrast, implicit queries require commonsense knowledge; for example, processing "an escape direction in case of an earthquake" requires the model to infer safety by identifying open areas and avoiding dense structures.

Linguistic Variability. As shown in Figure 1, this incorporates both concise and long, descriptive queries to evaluate model robustness to varying levels of linguistic detail.

7.2. The Details of Masks Generation

As shown in Figure 2, we employ two distinct point prompting strategies to generate candidate masks. For global point prompts, we set a grid parameter $R = C = 4$, uniformly sampling an $R \times C$ grid of 16 points across the square image to prompt the SAM. For local point prompts, we first utilize the provided bbox annotation to crop the target region. Points are then sampled exclusively within this cropped area. Assuming the cropped region has a height of h and a width of w, the sampling grid dimensions (R, C) are determined by:

$$(R, C) = \begin{cases} \left(\left\lceil \frac{\max\{h,w\}}{\min\{h,w\}} \right\rceil + 1, 1\right), & \text{if } \frac{\max\{h,w\}}{\min\{h,w\}} \geq 2.5 \\ (4, 4), & \text{if } 1 \leq \frac{\max\{h,w\}}{\min\{h,w\}} < 2.5 \end{cases}$$

7.3. The Details of Masks Filtering

In this section, we detail the automated mask filtering pipeline, which is visually depicted in Figure 2. This process is designed to programmatically curate the dataset by removing low-quality or erroneous annotations. We use the "airplane" category as an illustrative example to describe this multi-stage procedure.

First, we perform a coarse-grained sanity check based on object counts. We iterate through all samples and discard any instance where the number of discrete connected components in the binary mask does not precisely match the number of associated bounding boxes. This step effectively removes clear anomalies, such as fragmented masks (*i.e.*, multiple components for one box) or improperly merged masks (*i.e.*, one component for multiple boxes).

Second, from the pool of samples that pass this initial check, we establish a "gold standard" reference set. This set consists of 50 high-quality "airplane" masks that were manually selected and verified by human annotators. This reference set serves as the exemplar distribution from which we derive the target geometric profile for the category.

Third, we derive and apply a set of heuristic thresholds by statistically analyzing this "gold standard" set. We calculate the following geometric properties for the primary connected component of each of the 50 reference masks. All calculations are implemented using the OpenCV library.

- **Eccentricity**: Measures how much the shape of the mask deviates from a perfect circle.
- **Circularity**: Quantifies the "roundness" of the mask, defined as $\frac{4\pi \cdot \text{Area}}{\text{Perimeter}^2}$.
- **Solidity**: The ratio of the mask's area to the area of its convex hull. This metric penalizes shapes with significant indentations.
- **Symmetry**: A custom metric defined to quantify the object's expected bilateral symmetry.
- **Extent**: The ratio of the mask's area to the area of its bounding box ($\frac{\text{Mask Area}}{\text{Bbox Area}}$). This filters masks that are

overly sparse.

By calculating the distribution (*e.g.*, mean and standard deviation) of these properties across the reference set, we establish an "acceptable range" for each metric. Finally, these ranges are applied as a fine-grained filter to all remaining masks in the dataset. A mask is retained only if all its geometric properties fall within these predefined bounds.

7.4. The Details of QAs Generation

In this section, we detail the QA generation process, as shown in Figure 2. Our methodology adapts based on the mask source, which falls into two categories: those with pre-existing category labels and those without.

For masks with category labels, we employ two prompts, Prompt 6 and Prompt 7. Simple adjustments to these prompts generate a diverse set of QA pairs, allowing us to control factors such as output length and implicit reasoning.

Conversely, for masks that lack category labels, we first perform a preliminary category generation step. We use Prompt 8 to assign a category label to the masked region. Once the label is obtained, we then proceed to generate the corresponding QA pairs using the methods described above.

To execute this entire generation pipeline, we utilize the Gemini-2.5-Pro model, which we selected for its exceptional instruction-following capabilities. To ensure a high degree of diversity in the generated questions, we set the temperature parameter to 1.0. Furthermore, we enable its chain-of-thought reasoning capabilities to maintain high fidelity and coherence in the output.

7.5. The Details of QAs Filtering

The initial generation phases produced a raw dataset of approximately 40k QA pairs. To ensure the highest standards of quality and eliminate any generation artifacts, we implemented a rigorous, multi-stage manual validation process. This critical task was performed by a team of 15 domain experts, all with extensive experience in remote sensing and computer vision.

Each of the 40k QA pairs was meticulously inspected against a four-point quality rubric. A pair was only retained if it passed all four criteria:

- **Object recognition:** Verifying the accuracy of object category recognition in the image.
- **Spatial description and logical consistency of the question:** Ensuring the spatial and logical accuracy of the question's description.
- **Mask quality:** Assessing if the annotated mask meets quality standards for precision and completeness.
- **Grammatical accuracy:** Checking for grammatical mistakes or inconsistencies in the text.

Pairs that failed any of these checks were discarded. Following this comprehensive review, we performed a final quality assurance pass that involved randomly sampling the filtered set to verify the consistency and quality of the experts' work.

This expert-driven curation process yielded our final dataset, comprising 30,830 high-quality QA pairs that, in total, correspond to over 40,000 validated object masks.

7.6. Category List of LaSeRS

Table 7 illustrates the category distribution of the LaSeRS dataset. Our dataset includes not only general categories but also fine-grained concepts and part-level categories.

7.7. More Examples of LaSeRS

Figures 9 and 10 show more examples of LaSeRS.

8. Experiment Setting Details

8.1. Training Details

Table 8 presents the specific training hyperparameters of SegEarth-R2. Training and testing were conducted on an Nvidia A100 80GB GPU.

9. Visual Analysis

9.1. Visual Analysis of Spatial Attention

Figure 12 shows additional visualizations of the spatial attention mechanism. It can be observed that our spatial attention supervision helps the model better localize complex or small objects.

9.2. Visual Analysis of Segmentation Query

Figure 13, Figure 14, Figure 15 and Figure 16 show the ground truth mask, the mask selected by matching algorithm and the candidate masks when segmentation query number is set to 100. In many cases, the mask selected by the matching algorithm is not optimal, which leads to losses in both efficiency and performance. We set the segmentation query number to 1, and Figure 5 demonstrates the results. It can be observed that both efficiency and performance have been significantly improved.

10. Qualitative Results of SegEarth-R2

10.1. LaSeRS

Figure 17 and Figure 10 show the qualitative results on the LaSeRS test set.

10.2. RefSegRS, RRSIS-D, and RISBench

Figure 19, Figure 20, and Figure 21 show the qualitative results on the RefSegRS, RRSIS-D, and RISBench benchmarks, respectively.

10.3. EarthReason

Figure 22 show the qualitative results on EarthReason benchmark.

Table 7. Comparison of dataset categories: LaSeRS contains 122 categories, encompassing general, fine-grained, and part-level classes, and covering a wide range of remote sensing scenes such as land cover, vehicles, and natural landscapes. In contrast, RefSegRS, RRSIS-D, RISBench, and EarthReason contain only a dozen to around twenty categories.

Dataset	Categories
LaSeRS	**General**: *airplane, airport, airport runway, bare land, baseball diamond, baseball field, basketball court, beach, bridge, bridge road, building, bushes, canal, chimney, cooling tower, dam, expressway service area, expressway toll station, farmland, football field, golf field, grass, green strip, greenhouse, ground track field, harbor, helicopter, helipad, intersection, jet bridge, lake, large vehicle, overpass, parking lot, path, paved road, paved square, plane, playground, railway, river, road, roundabout, sea, ship, slide, small car, small vehicle, soccer ball field, solar panel, sports field, stadium, storage tank, substation, swimming pool, tennis court, terminal, train station, tree, unimproved road, vehicle, volleyball court, water, white smoke, windmill.* **Fine-grained Concept**: *B1-B boomber, a220, a321, a330, a350, arj21, boeing737, boeing747, boeing777, boeing787, bus, c919, cargo truck, container crane, dry cargo ship, dockside warehouse, dump truck, driveway, engineering ship, excavator, fishing boat, hangar, liquid cargo ship, motorboat, passenger ship, tractor, trailer, truck tractor, tugboat, van, warship.* ***Part****: airplane engine, bleachers, bow of ship, cargo hold, center circle, center line, center service line, cooling tower shell, cooling tower top opening, downstream, football net, fuselage, horizontal stabilizer, industrial pipeline, net, no man's land of tennis court, riverbank, service box, shipping container, stern of ship, tennis net, three-point line, upstream, wake, wing, zebra crossing.*
RefSegRS [72]	**General**: *road, vehicle, car, van, building, truck, trailer, bus, road marking, bikeway, sidewalk, tree, low vegetation, impervious surface.*
RRSIS-D [36]	**General**: *airplane, airport, golf field, expressway service area, baseball field, stadium, ground track field, storage tank, basketball court, chimney, tennis court, overpass, train station, ship, express toll station, dam, harbor, bridge, vehicle, windmill.*
RISBench [9]	**General**: *expressive service area, expressive toll station, ground track field, basketball court, container crane, roundabout, windmill, overpass, stadium, bridge, soccer ball field, baseball diamond, train station, golf field, airport, harbor, dam, ship, helipad, vehicle, chimney, airplane, helicopter, tennis court, storage tank, swimming pool.*
EarthReason [28]	**General**: *storage tank, bridge, intersection, tennis court, baseball field, substation, pier, viaduct, wind turbine, church, airport runway, swimming pool, lake, airport helipad, dam, railway, basketball court, beach, greenhouse, roundabout, solar power plant, ground track field, waterwaste plant, river, train station, stadium, island, factory.*

Table 8. Hyper parameters of our model in the training.

Parameters	**Value**
Optimizer	AdamW
Learning Rate	1×10^{-4}
Batch Size	4
Training Steps	50,000
Learning Rate Schedule	Cosine Decay
Lora Rank	8
Computation Precision	bf16
Weight Decay	0.0
Warmup Ratio	0.03
Image Size	1024×1024
Image Processing	Resize long edge to 1024 and padding short edge to 1024.

Prompt:

Assume you are an expert in interpreting remote sensing imagery. I will provide you with a remote sensing image with annotations, where the annotated area is a target region outlined in red. I will also provide the object category name corresponding to this region. Based on the natural and socio-economic environment in the image, generate one implicit reasoning question-answer pair. Follow these requirements strictly:

1. The question must not directly mention the object category name. It should be implicit, requiring reasoning to infer the object type of the target region.
2. Imagine a realistic scenario based on the image context and describe the question in detail with complexity. Limit the question length between 15 and 40 words.
3. The question should have practical value in real life, such as guiding people's production or daily activities.
4. The question must reflect the distinctive characteristics of the target object type, and should not be easily confused with other objects in the image.
5. Generate a reasonable answer to the question, and the answer must explicitly include the object category name I provided.
6. Output in strict JSON format, and no line breaks or other characters are allowed inside the curly braces: {"description": your question, "answer": your answer}.
7. The generated questions and answers must not include phrases such as 'outlined in red'.
8. Note: If the image contains annotations for multiple red-outlined areas, the plural form of the category name should be used in the answer.
9. For specific types of aircraft or vehicles, their locations can be roughly indicated, such as in the lower right part of the image.

Here are several examples:
1. {"description": "In this industrial area near the river, exhaust gases or smoke generated during industrial processes are discharged into the upper atmosphere to dilute and disperse pollutants, thereby reducing their impact on the ground-level environment and human health.", "answer": "The chimney near the river can serve this purpose."}
2. {"description": "In this ocean area, if I want to find a small watercraft powered by jet propulsion that can glide rapidly on the water surface for recreational purposes, where should I look?", "answer": "The motorboat in the lower left part of the image meets your requirements."}
3. {"description": "In this busy residential area, where roads crisscross, which locations are prone to traffic accidents and therefore suitable for installing traffic lights and pedestrian crossings?", "answer": "The intersection in the image is a high-risk area for traffic accidents."}
4. {"description": "In this airport, which aircraft in the lower-left of image is the smaller-sized one equipped with the Pratt & Whitney PW1500G geared turbofan engine?", "answer": "The a220 aircraft is a smaller-sized plane that is equipped with the Pratt & Whitney PW1500G geared turbofan engine."}
5. {"description": "Which area in this image is suitable for 5-on-5 team sports that emphasize teamwork and coordination, where players need to perform skilled passing, screening, and movement?", "answer": "The basketball court in the lower right part of the image is suitable for 5-on-5 team sports."}

The image with red contours is provided below. The category name of target region is ***{category_name}***. Please analyze the image and generate a question-answer pair based on the above requirements.

Figure 6. An example of single-target QA generation prompt

Prompt:

Assume you are an expert in remote sensing interpretation. I will provide you with a remote sensing image annotated with multiple colored outline boxes, where each color highlights a region containing a specific land cover or object type. I will also provide you with the object type represented within each color outline. Your task is to generate a pair of implicit reasoning-based question and answer according to the following rules:

1. The question must be *implicit*, meaning it should not directly mention any of the provided category names. Instead, the question should require reasoning to infer the object types present in the different outlined regions.
2. You may reasonably assume or imagine a scene to construct an integrated question, or construct individual sub-questions based on each color's outline and connect them using 'and'.
3. The question must reflect *distinct characteristics* of the object types in each colored region.
4. You must generate a *plausible answer* to the question. The answer must include *all* of the given category names.
5. Format the output as a strict JSON string with no extra characters or line breaks inside the brackets: {"description": your question, "answer": your answer}
6. Do *not* use phrases like "outlined in red" or "outlined in blue" in the answer.
7. If a certain category appears in *multiple instances* in the image, use the *plural* form of that category name in the answer.
8. You may refer to spatial positions such as "in the center of the image" or "in the lower-right corner" in the question to help structure the reasoning.
9. Leverage your *common sense knowledge and creativity* to generate the implicit reasoning question. The final question must implicitly involve all the categories I provide.

Below are several examples for your reference:
1. Given categories: small car, dump truck, storage tank
Output: {"description": "Which vehicles in the image are used to transport a small number of people or for self-driving tours, which ones are primarily used to carry soil or rocks, and which large structure might be used to store water or industrial materials?", "answer": "The small cars scattered across various corners of the image are suitable for personal travel or self-driving tours, the dump truck in the bottom right is used for transporting soil and rocks, and the large storage tank in the center is a facility for storing industrial materials."}
2. Given categories: basketball court, tennis court, baseball field
Output: {"description": "In this sports complex, three types of sports competitions are taking place: a game where the ball is thrown into a hoop, a rally sport involving hitting a ball over a net, and a sport where the ball is struck within a fan-shaped area. Please identify the areas where each of these sports is being played.", "answer": "The basketball court is the area for shooting the ball, the tennis court is the area for hitting the ball over the net, and the baseball field is the fan-shaped area."}
3. Given categories: a220, a330, boeing737
Output: {"description": "In this airport area, which two types of aircraft are manufactured by the European company Airbus—one being a single-aisle small passenger jet and the other a twin-aisle medium-sized aircraft—and which popular aircraft is produced by the American company Boeing?", "answer": "The a220s are small aircraft produced by Airbus, the a330s are medium-sized aircraft from Airbus, and the boeing737s is a popular passenger jet manufactured in the United States."}

The red outlined areas in the image represent ***{category_name_1}***, the blue outlined areas represent ***{category_name_2}***, the orange-yellow outlined areas represent *{category_name_3}*. Please follow the above rules to generate the question-answer pair.

Figure 7. An example of multi-target QA generation prompt

Categories list:

["parking lot", "tree", "canal", "river", "farmland", "cooling tower shell", "cooling tower top opening", "industrial pipeline", "B1-B boomber","airplane","building", "terminal", "jet bridge", "sports field", "basketball court","tennis court", "unimproved road", "center service line", "service box", "path","baseball field", "center circle", "three-point line", "airplane engine", "hangar","lake", "upstream", "paved square", "overpass", "sea", "storage tank", "bleachers", "ship", "harbor", "shipping container", "helipad", "airport runway", "white smoke", “dam”, “driveway”, “football field”, “bushes”, “riverbank”, “road”, “vehicle”, “railway”, “airport”, "cargo hold", "volleyball court", "roundabout", "airport runway", "swimming pool", "solar panel", "dockside warehouse", "wake", "golf field", "beach", "substation", "bow of ship", "stern of ship", "zebra crossing", "no man's land of tennis court", "intersection"]

Prompt:

Assuming you are an expert in remote sensing image interpretation, I will provide a remote sensing image that contains multiple closed regions enclosed by red outlines. Additionally, I will provide a list of region categories. The categories list is ***{categories_list}***. you should select a category from the list that corresponds to a red-outlined region according to the following requirements:

1. The final output must be only the object's category name. Do not include any other redundant output, such as 'the red-outlined region is...'
2. The output category name must be selected from the categories list.
3. Pay attention to the differences between similar categories, such as cooling tower shell and cooling tower top opening, path, road and road bridge, tree and bushes.
4. Before submitting your final result, please carefully verify that your category selection accurately reflects the red-outlined region and avoids any ambiguity.

Examples:

1. If the image shows a single airplane on an airport taxiway, outlined in red. -> Output: airplane
2. If the image shows a residential area with one red contour drawn around three separate houses. -> Output: house
3. If the image shows a seaside port, and the red contour contains a warehouse-like building next to it. -> Output: dockside warehouse

Figure 8. Category generation prompt

QA

Q: Can you locate the cooling tower shell of the upper cooling tower.

A: Of course! The *<p> cooling tower shell </p> [SEG]* segmentation completed.

Q: Which component of the large airplane at the bottom is crucial for maintaining the longitudinal stability of flight?

A: The ***<p> horizontal stabilizer <p> [SEG]*** is crucial for maintaining the longitudinal stability of flight. It helps control the aircraft's pitch and prevents unwanted nose-up or nose-down movements, ensuring stable flight.

Q: Which part of the water body in this area has unique conditions suitable for power generation and plays a significant role in the economic development of the surrounding regions?

A: The area *<p> upstream </p> [SEG]* of the river is an ideal location for building a hydroelectric power station.

Q: Can you locate the large, white airport terminal on the right, where passengers board and deplane aircraft, as a central hub for airport operations.

A: Done. The *<p> terminal </p> [SEG]* is now segmented.

Q: On a hot summer day, when you're sweating profusely, which spot in the image would be a good place to cool off?

A: The shade created by the *<p> tree </p> [SEG]* to the bottom right of the parking lot in the image is a great spot to cool off.

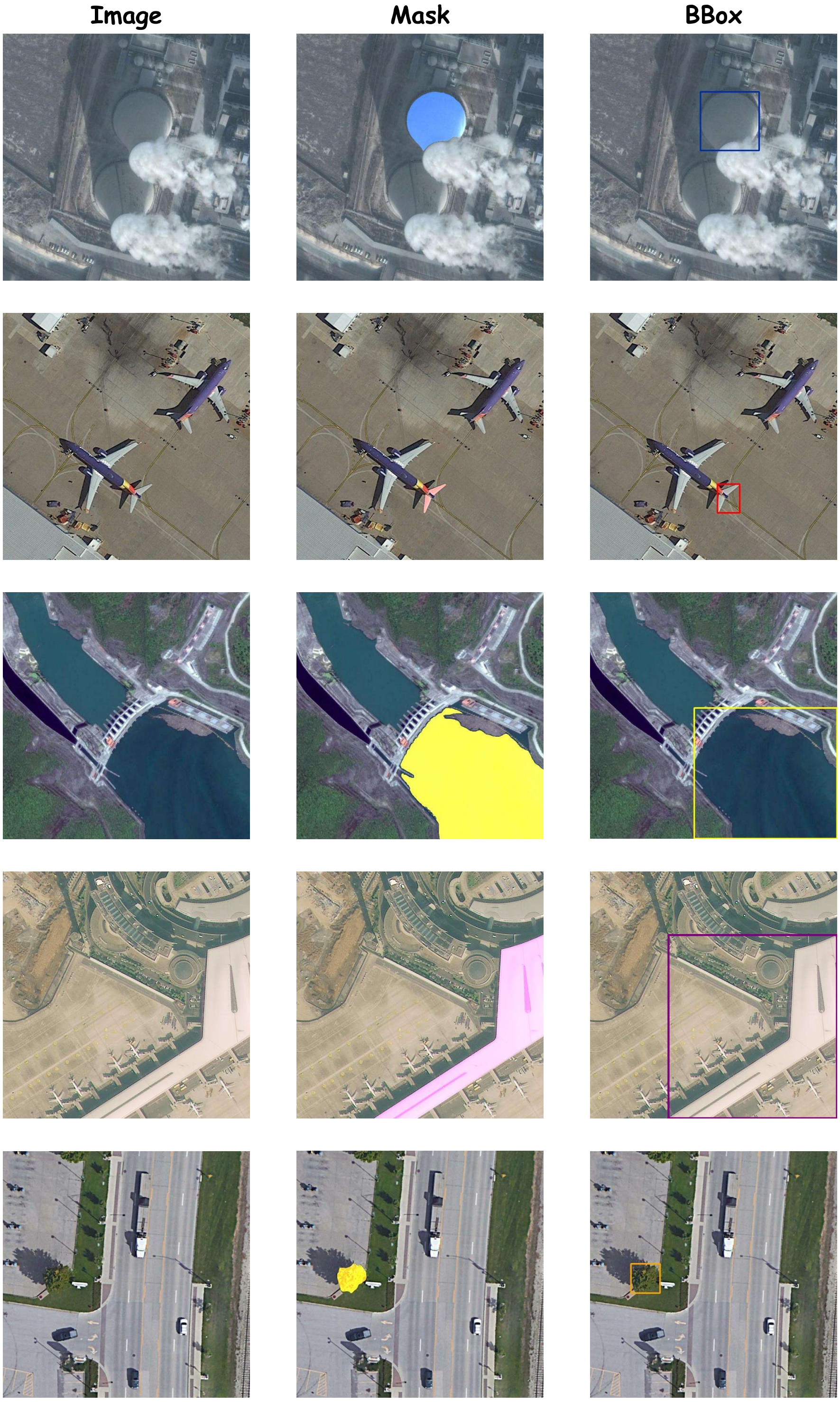

Figure 9. More examples of LaSeRS.

QA

Q: A logistics company is managing operations across these two lots. Which vehicle in the bottom yard is primarily used as the motive power unit for a semi-trailer, and which vehicle in the top parking lot is a self-contained unit used for local deliveries?

A: The *<p> truck tractor </p> [SEG]* is the motive power unit for a semi-trailer, and the *<p> cargo truck </p> [SEG]* is the self-contained unit for local deliveries.

Q: In this urban landscape, which major transportation hub is designed for the arrival and departure of trains, and which facility in the upper-left corner is specifically built for athletic events like track and field competitions?

A: The *<p> train station </p> [SEG]* is the transportation hub for trains, and the ***<p> ground track field </p> [SEG]*** is the facility for athletic events.

Q: In this transportation hub, which vehicles parked in the central lot are designed for mass transit, capable of carrying numerous passengers, and which vehicles on the roads are typically used for personal or family travel with a much smaller capacity?

A: The *<p> large vehicles </p> [SEG]* are designed for mass transit, and the *<p> small vehicles </p> [SEG]* are used for personal travel.

Q: At this military airbase, which aircraft in the upper-left corner are designed for vertical takeoff and landing, often used for transport and reconnaissance, and which aircraft in the center are fixed-wing vehicles that require a runway and are typically used for high-speed combat missions?

A: The *<p> helicopters </p> [SEG]* are capable of vertical takeoff and landing, while the *<p> planes </p> [SEG]* are fixed-wing combat aircraft.

Q: At this airport, which aircraft is a famous four-engine model from a major American manufacturer? Which two are the world's largest twin-engine jets from the same company? And which two twin-engine wide-body aircraft are from its main European competitor?

A: The *<p> boeing747 </p> [SEG]* is the famous four-engine American model, the *<p> boeing777s </p> [SEG]* are the world's largest twin-engine jets from the same company, and the *<p> a330 </p> [SEG]* and *<p> a350 </p> [SEG]* are the twin-engine wide-body aircraft from its European competitor.

Figure 10. More examples of LaSeRS.

Figure 11. More examples of LaSeRS.

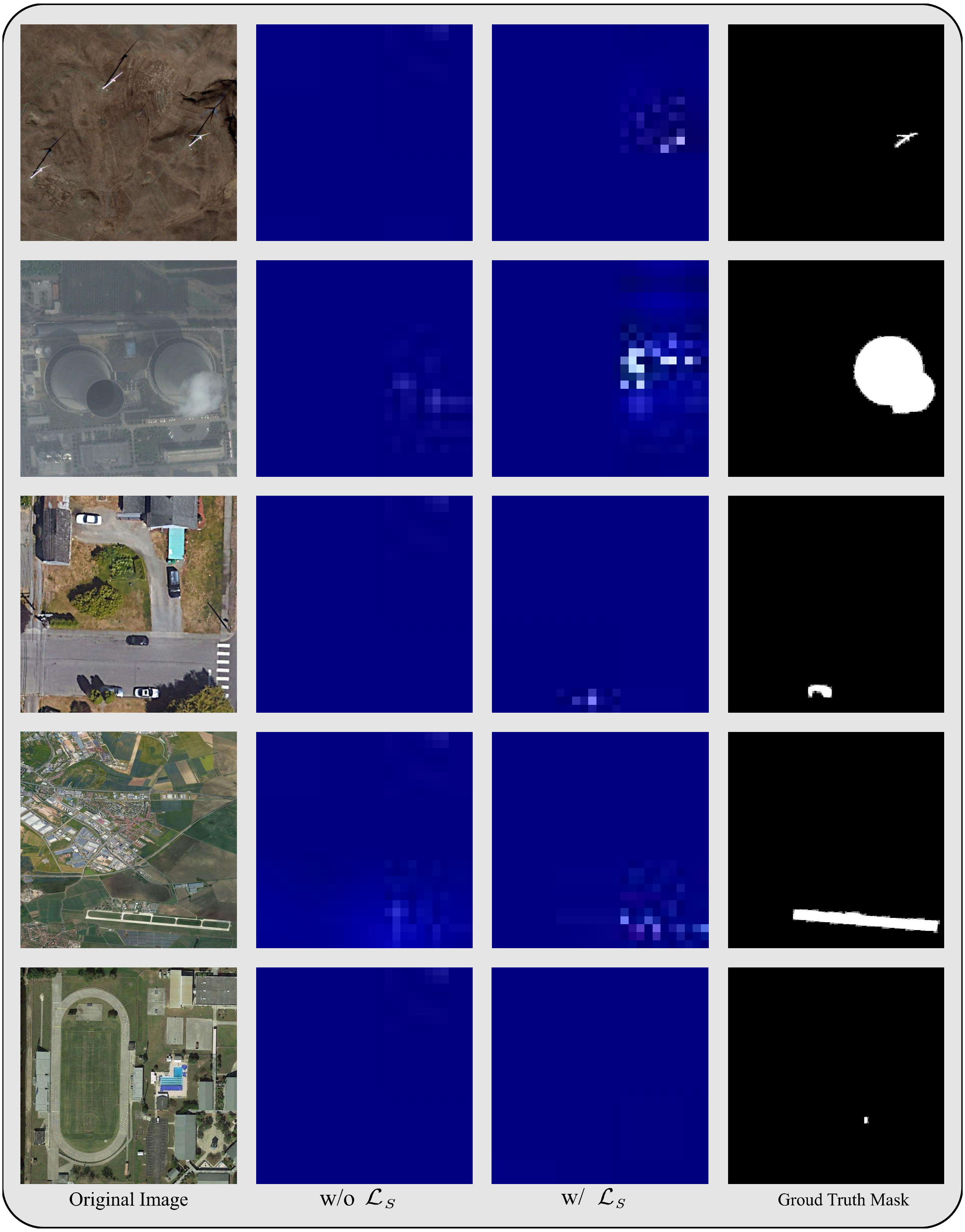

Figure 12. Additional examples of spatial attention visualizations.

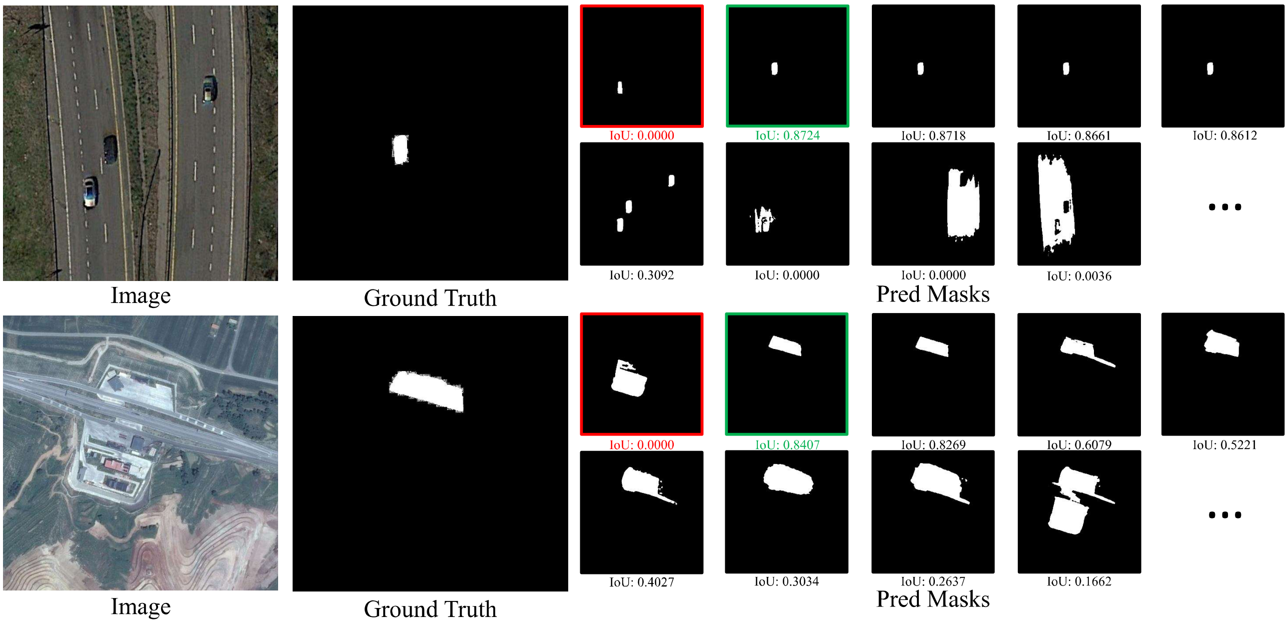

Figure 13. Visualization analysis of segmentation query number. The red box indicates the mask selected by the matching algorithm, while the green shows the mask that best matches the ground truth. It can be observed that the matching algorithm always selects a non-optimal mask, leading to a decrease in both efficiency and performance.

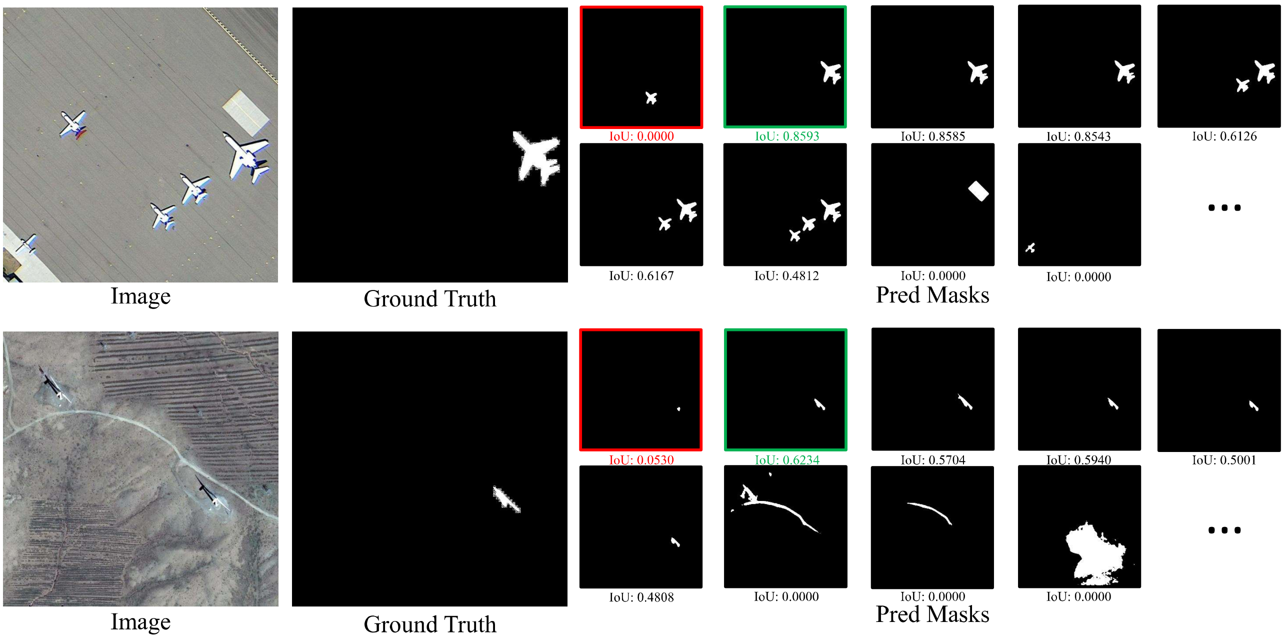

Figure 14. Visualization analysis of segmentation query number. The red box indicates the mask selected by the matching algorithm, while the green shows the mask that best matches the ground truth. It can be observed that the matching algorithm always selects a non-optimal mask, leading to a decrease in both efficiency and performance.

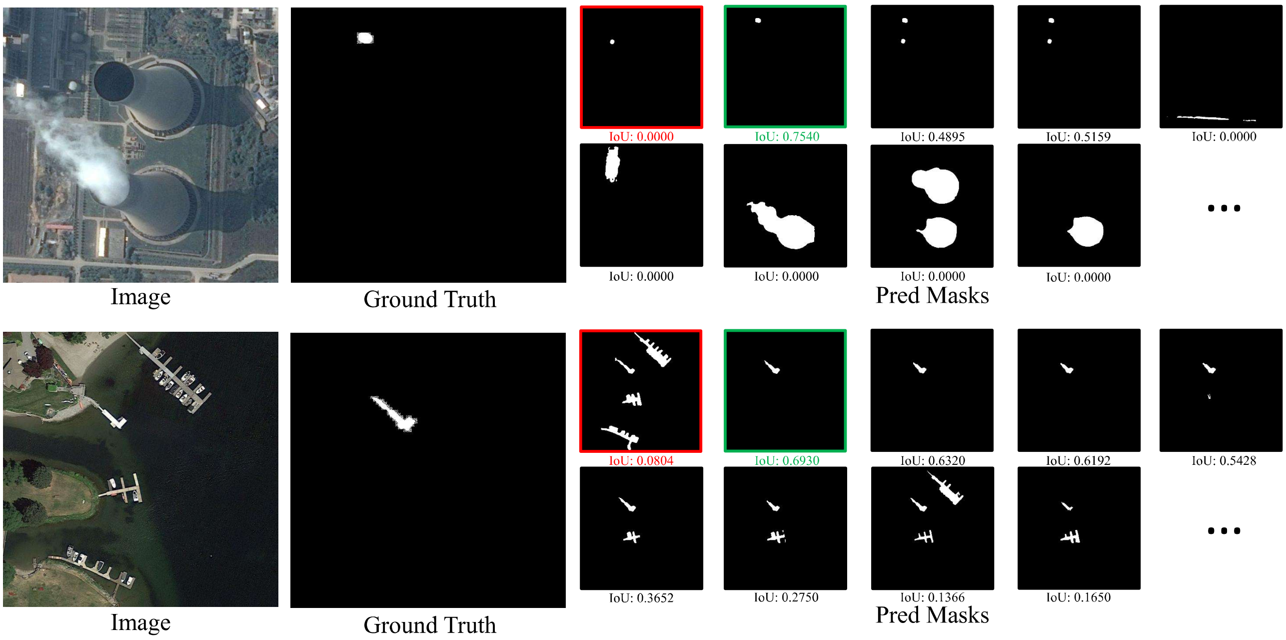

Figure 15. Visualization analysis of segmentation query number. The red box indicates the mask selected by the matching algorithm, while the green shows the mask that best matches the ground truth. It can be observed that the matching algorithm always selects a non-optimal mask, leading to a decrease in both efficiency and performance.

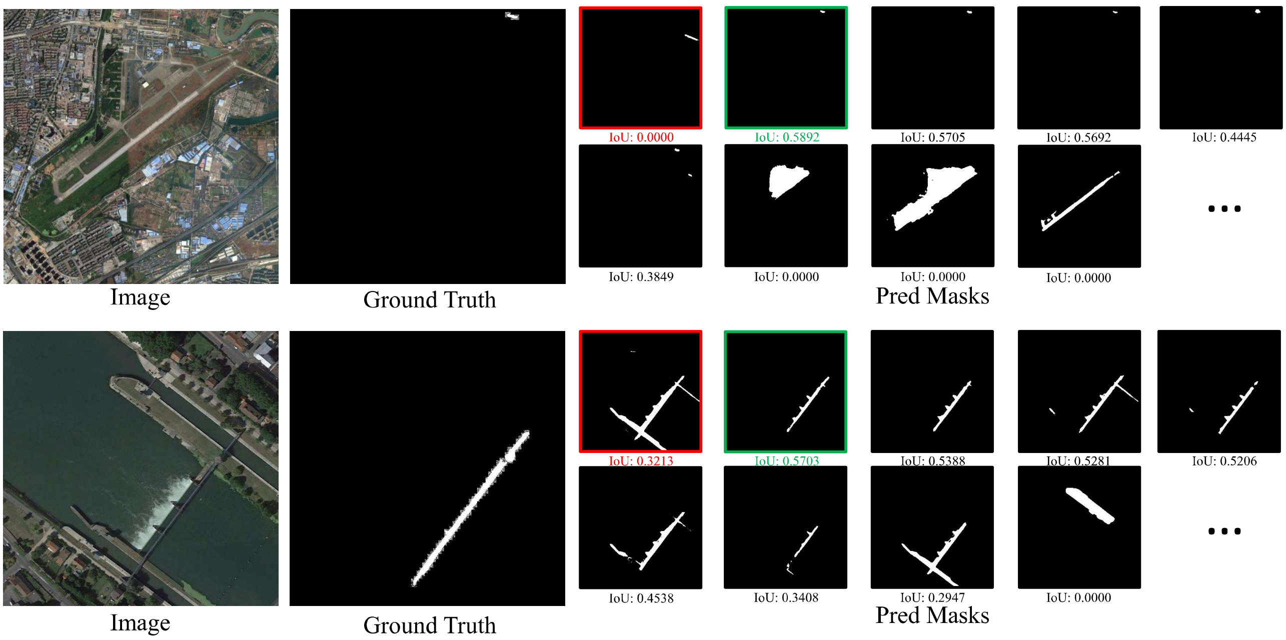

Figure 16. Visualization analysis of segmentation query number. The red box indicates the mask selected by the matching algorithm, while the green shows the mask that best matches the ground truth. It can be observed that the matching algorithm always selects a non-optimal mask, leading to a decrease in both efficiency and performance.

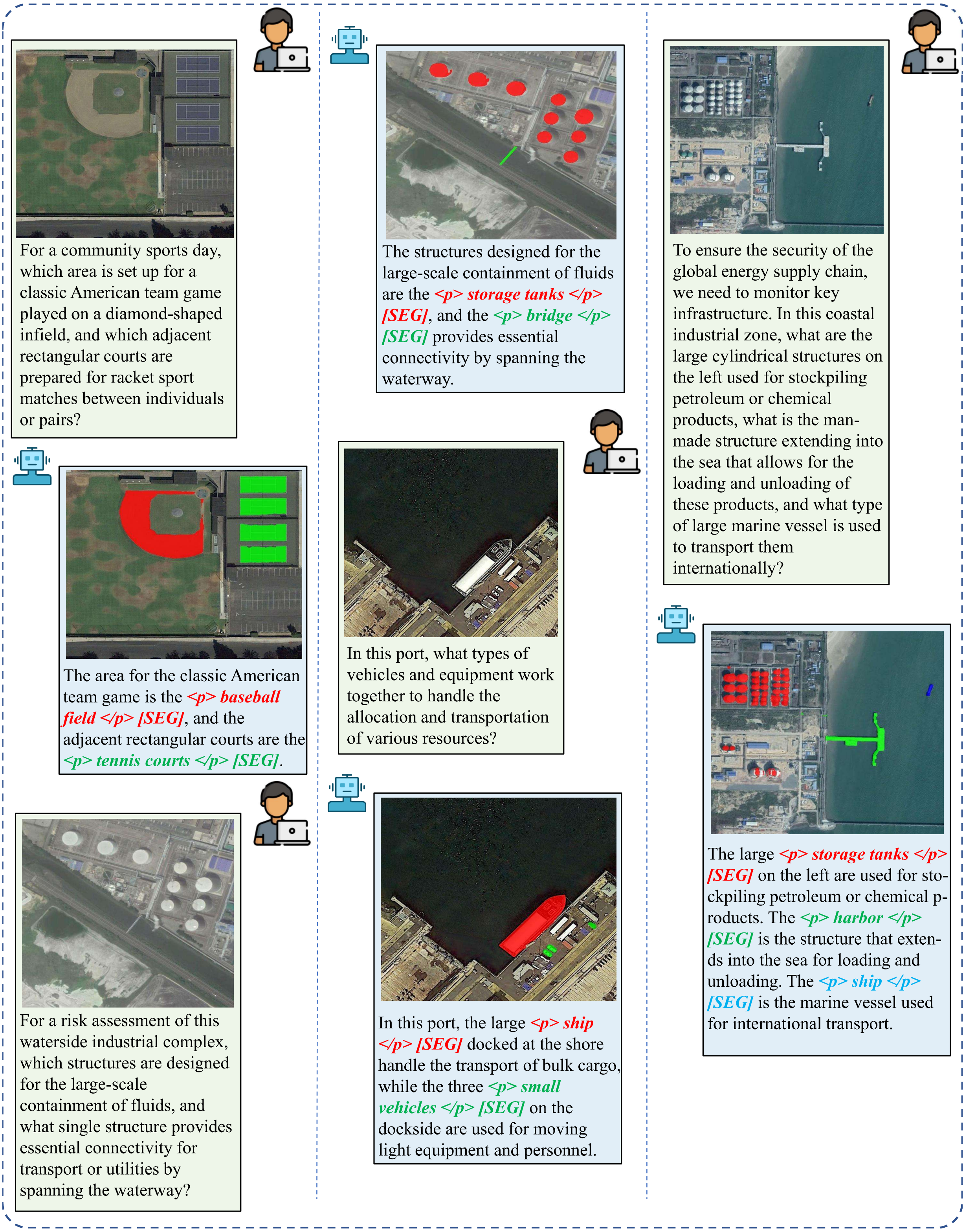

Figure 17. Visualization results of SegEarth-R2 on the LaSeRS dataset.

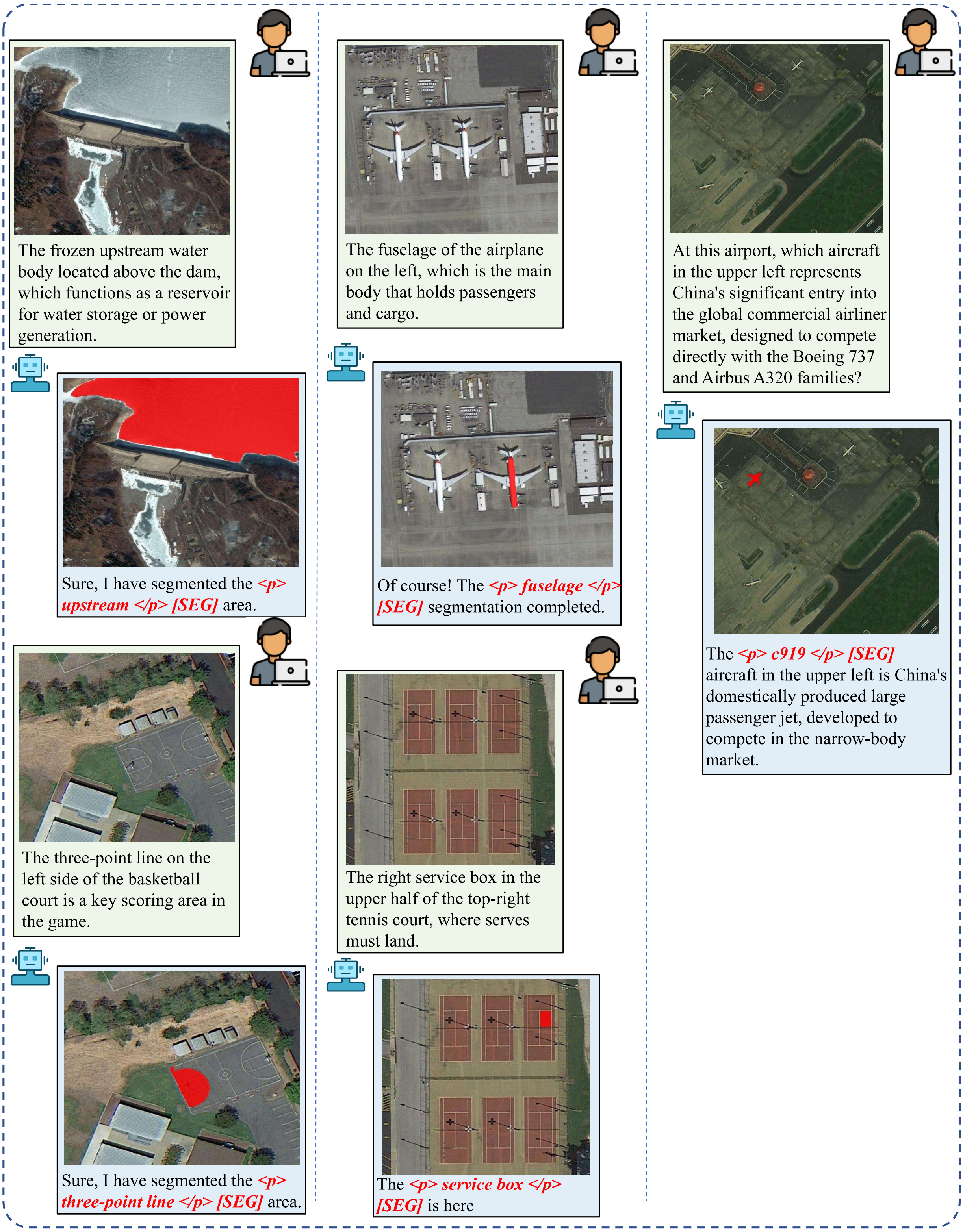

Figure 18. Visualization results of SegEarth-R2 on the LaSeRS dataset.

Figure 19. Visualization results of SegEarth-R2 on the RefSegRS test set.

Figure 20. Visualization results of SegEarth-R2 on the RRSIS-D test set.

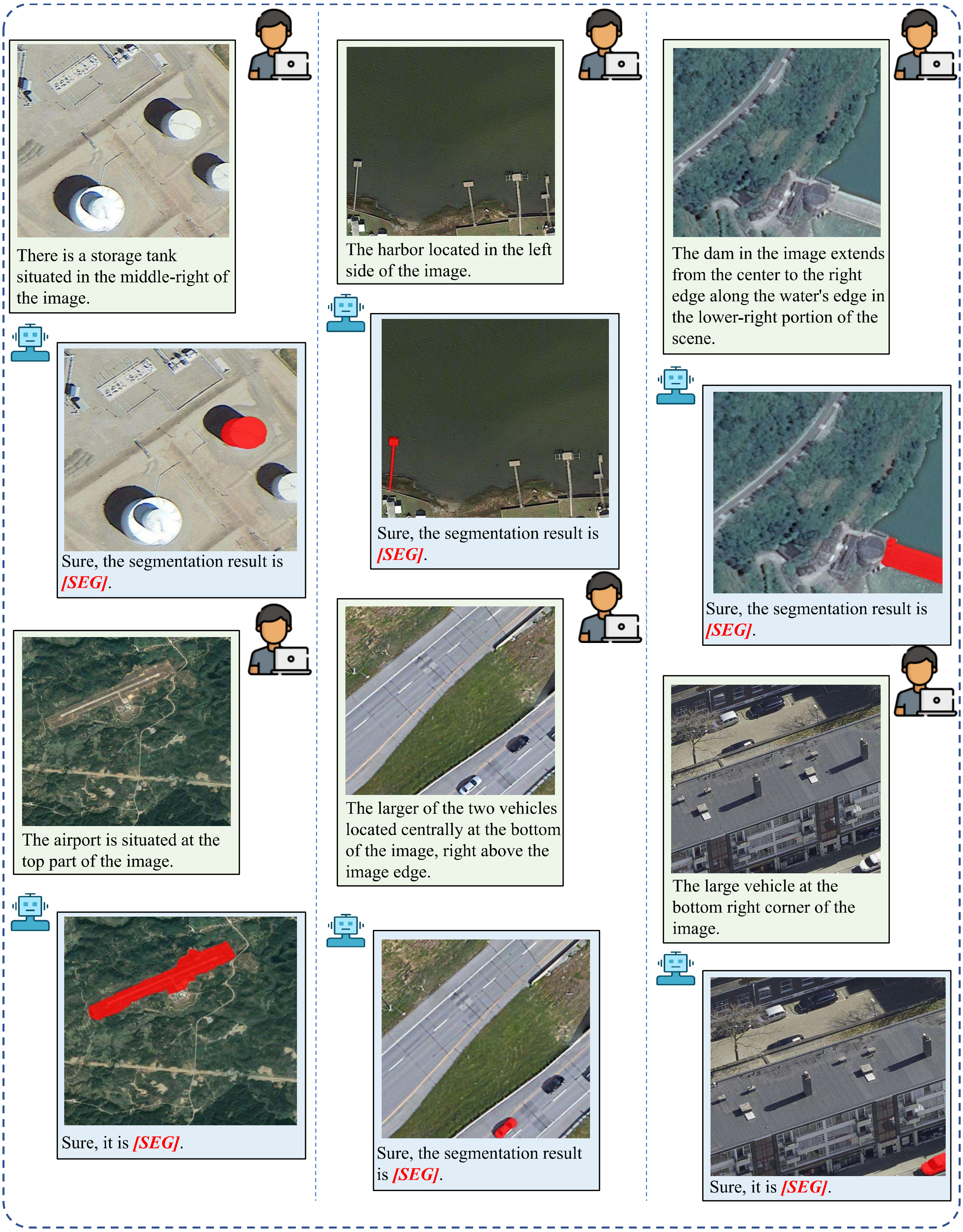

Figure 21. Visualization results of SegEarth-R2 on the RISBench test set.

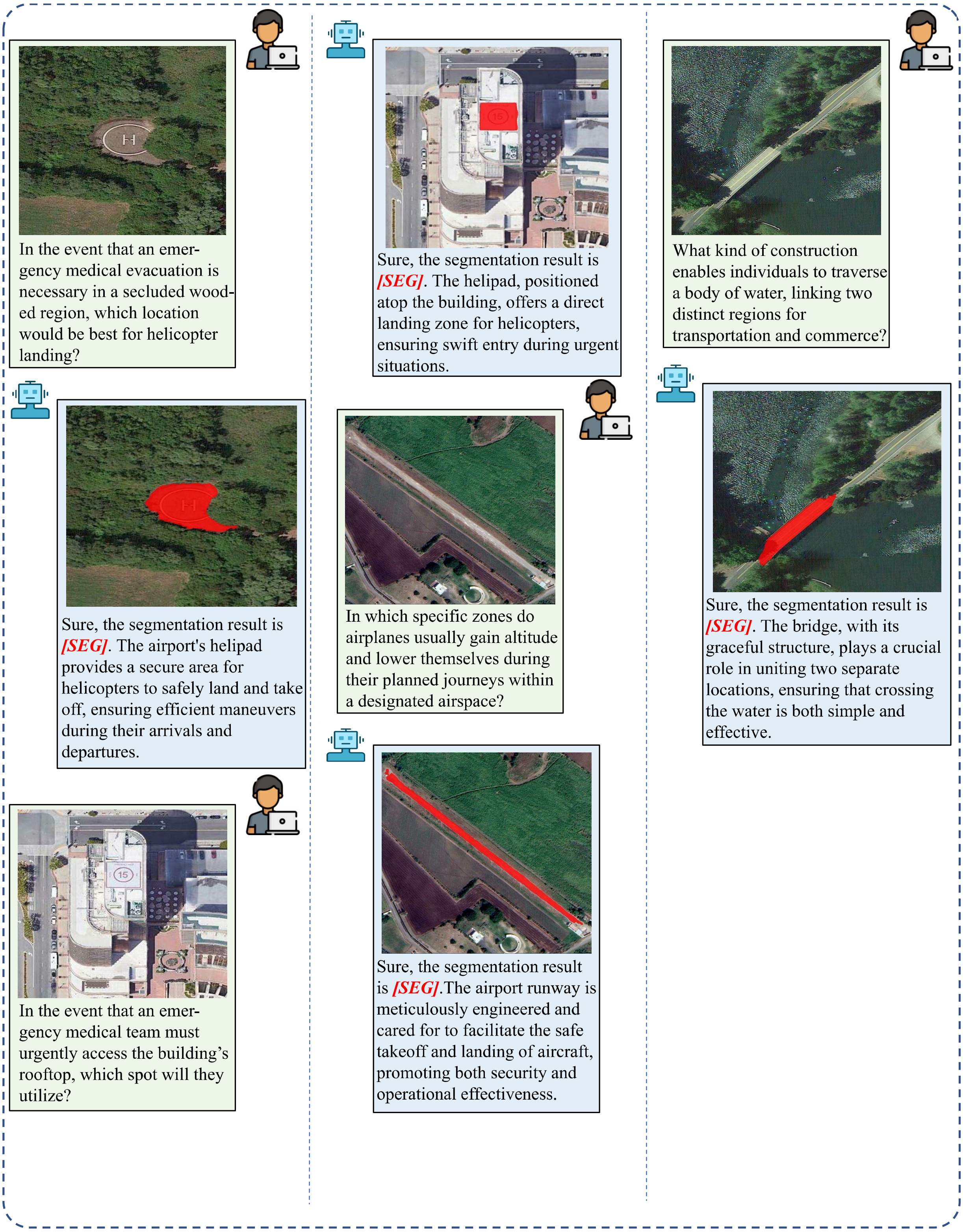

Figure 22. Visualization results of SegEarth-R2 on the EarthReason test set.