1 Overview of the Appendix

This appendix supplements the proposed **XLRS-Bench** with additional experimental results and details excluded from the main paper due to space constraints.

The appendix is organized as follows:

- Sec. 2: More details of XLRS-Bench.
- Sec. 3: Human evaluations on XLRS-Bench.
- Sec. 4: More analysis on L-2 capability across various MLLMs.
- Sec. 5: Detailed results of specific sub-tasks (L-3 capability).
- Sec. 6: Visualizations of samples and challenging cases.
- Sec. 7: Datasheets for the XLRS-Bench dataset.
- Sec. 8: Discussion on limitations and societal impact.

2 More Details of XLRS-Bench

Table 1: Characterristics and vision-language formats of L3 sub-tasks.

L1-Task	L2-Task	L3-Task	Abbr.	Annotation Format	Annotation Method	Number of Samples	Answer Type
Perception	Counting	Overall Counting	OC	VQA	All Human	370	Multiple Choice(A/B/C/D)
		Regional Counting	RC	VQA	All Human	972	Multiple Choice(A/B/C/D)
	Scene Classification	Overall Land Use Classification	OLUC	VQA	All Human	904	Multiple Choice(A/B/C/D)
		Regional Land Use Classification	RLUC	VQA	All Human	1854	Multiple Choice(A/B/C/D)
	Object Spatial Relationship	Object Spatial Relationship	OSR	VQA	All Human	4819	Multiple Choice(A/B/C/D)
	Object Properties	Object Classification	OCC	VQA	All Human	9172	Multiple Choice(A/B/C/D)
		Object Color	OCL	VQA	All Human	8930	Multiple Choice(A/B/C/D)
		Object Motion State	OMS	VQA	All Human	640	Multiple Choice(A/B for Yes/No)
	Image Captioning	Detailed Image Captioning	-	Caption	Semi-automated	934	Plain Text
	Visual Grounding	Fine-grained Visual Grounding	-	Visual Grounding	All Human	6310	Bounding Box
Reasoning		Condition-based Visual Grounding	-	Visual Grounding	All Human	6305	Bounding Box
	Route Planning	Route Planning	RP	VQA	All Human	1130	Multiple Choice(A/B/C/D)
	Anomaly Reasoning	Anomaly Detection and Interpretation	AD	VQA	All Human	1131	Multiple Choice(A/B/C/D)
	Complex Reasoning	Environmental Condition Reasoning	ECR	VQA	All Human	1125	Multiple Choice(A/B/C/D)
		Counting with Complex Reasoning	CCR	VQA	All Human	972	Multiple Choice(A/B/C/D)
	Spatiotemporal Reasoning	Regional Counting with Change Detection	RCCD	VQA	All Human	270	Multiple Choice(A/B/C/D)

We provide additional details about the dataset, with Table 1 presenting statistics for VQA, visual grounding, and image captioning tasks, along with their relationships to the L3 sub-tasks. This clarifies the dataset's structure and composition. Notably, Visual Grounding spans both perception and reasoning, with Fine-grained Visual Grounding classified under perception and Condition-based Visual Grounding under reasoning.

3 Human Evaluations on XLRS-Bench

Human evaluation is essential for assessing dataset effectiveness [1]. For XLRS-Bench, we randomly selected 30 questions from each VQA sub-task (L-3 dimensions) and had two groups answer them simultaneously. The final accuracy was computed as the average accuracy of both groups. Figure 1 illustrates the evaluation results of MLLMs and humans.

We observed that human accuracy consistently exceeded 90%, validating the reliability of XLRS-Bench. However, human evaluation is not error-free, as analyzing large ultra-high-resolution RS images demands intense focus and frequent zooming, particularly for tasks like global counting, making it inherently challenging. In contrast, existing MLLMs, such as the closed-source GPT-4o, performed significantly worse, likely due to insufficient training on real ultra-high-resolution RS data. We encourage future research to address these challenges.

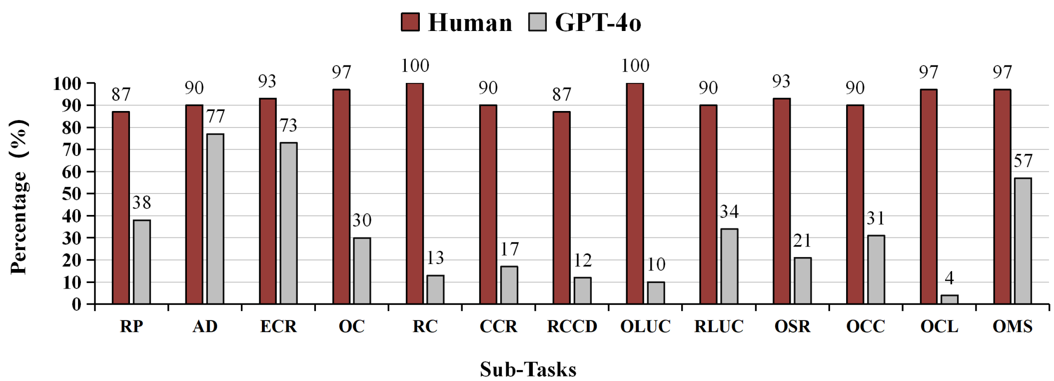


Figure 1: **Evaluation results of XLRS-Bench and MLLMs.** “RP”, “AD”, “ECR”, “OCC”, “RC”, “CCR”, “RCCD”, “OLUC”, “RLUC”, “OSR”, “OCC”, “OCL”, and “OMS” each indicate a specific task domain: Route Planning, Anomaly Detection, Environmental Conditional Reasoning, Overall Counting, Regional Counting, Counting with Complex Reasoning, Regional Counting with Change Detection, Overall Land Use Classification, Regional Land Use Classification, Object Spatial Relationship, Object Classification, Object Color and Object Motion State.

4 More Analysis of Results on XLRS-Bench

Due to space limitations, more in-depth analyses to advance MLLM research in ultra-high-resolution remote sensing scenarios are provided in the appendix. This section highlights the performance of all L-2 capabilities.

Most MLLMs underperform across all 16 evaluation dimensions. The accuracy of most MLLMs remains below 50%, in sharp contrast to the 80%–90% typically observed in common benchmarks [1, 2, 3, 4]. Notably, the high accuracy and minimal variation among advanced models in these benchmarks often obscure their practical utility, reducing the significance of small improvements. The consistently low performance on XLRS-Bench underscores the distinct challenges of ultra-high-resolution remote sensing, driven by a lack of pretraining on annotated data. This highlights the pressing need for specialized models to address these complexities.

Performance Gap: Anomaly vs. Spatiotemporal Reasoning A notable performance gap exists between Anomaly Reasoning (AR) and Spatiotemporal Reasoning (SR) tasks. While most models achieve about 70% accuracy on AR tasks, their performance drops sharply to 15.2% on SR tasks. This discrepancy arises because AR tasks depend on identifying global anomalies with clear patterns, whereas SR tasks demand intricate local spatiotemporal modeling. Current MLLMs excel at detecting static anomalies but struggle with dynamic pattern comprehension. To bridge this gap, optimizing MLLMs should focus on improving temporal feature modeling, such as enhancing Transformer architectures to better handle sequential data.

Limited Benefits of Larger LLMs in Perception Tasks. In the Counting and Scene Classification (SC) subtasks, LLaVA-Next (Llama3-8B) offers little advantage over the smaller Qwen2-VL (Qwen2-7B), indicating that model size is not a primary determinant of performance. Instead, factors like diverse pretraining data and effective task alignment mechanisms likely play a more significant role. This underscores the reliance of perception tasks on the visual module’s capabilities rather than the language model’s reasoning. Future efforts could prioritize smaller, more efficient models tailored for

Table 2: **Experimental results of L-3 capability on the perception dimension of VQA tasks.** Models are ranked according to their average performance. Rows corresponding to proprietary models are highlighted in gray for distinction. "OC", "RC", "OLUC", "RLUC", "OSR","OCC","OCL" and "OMS" each indicate a specific task domain: Overall counting, Regional Counting, Overall Land Use classification, Regional Land Use Classification, Object Spatial Relationship, Object Classification, Object Color and Object Motion State.

Method	LLM	Language	Perception								
Subtasks (L-3 Capability)			OC	RC	OLUC	RLUC	OSR	OCC	OCL	OMS	Avg
CogVLM2	Llama3-8B	en	31.89	38.48	1.53	68.07	**35.92**	**40.77**	30.93	**64.53**	**37.65**
LLaVA-OneVision	Qwen2-7B	en	29.19	38.37	1.11	**70.50**	32.60	35.63	**34.37**	62.50	36.54
Qwen2-VL	Qwen2-7B	en	**32.43**	**42.49**	5.86	68.99	32.04	35.05	33.34	59.53	36.09
GPT-4o-mini	-	en	20.27	31.38	**18.69**	58.52	29.96	39.97	30.56	63.44	35.71
InternVL2	InternLM2.5-7B	en	22.97	38.07	8.19	62.84	26.60	35.02	32.79	60.94	34.37
InternLM-XComposer-2.5	InternLM2-7B	en	26.76	39.30	1.11	70.01	32.75	33.41	29.06	11.72	32.90
LLaVA-Next	Llama3-8B	en	27.84	41.87	1.11	60.14	32.50	29.60	30.59	63.12	32.72
GPT-4o	-	en	24.32	31.48	16.81	64.02	32.35	19.59	29.00	40.31	28.70
LLaVA-1.5	Vicuna-7B	en	24.05	22.12	0.00	29.29	23.05	21.65	17.45	38.12	20.77
GeoChat	Vicuna-7B	en	24.05	22.12	1.00	27.72	23.30	21.65	17.45	38.13	20.74
Qwen2-VL	Qwen2-7B	zh	**32.16**	42.28	1.77	**72.44**	33.26	**40.07**	**34.06**	60.00	**38.29**
InternVL2	InternLM2.5-7B	zh	22.43	38.99	3.65	56.42	34.24	39.60	33.75	60.78	36.97
LLaVA-OneVision	Qwen2-7B	zh	27.03	41.46	1.11	69.47	31.73	39.00	28.33	62.50	35.56
GPT-4o-mini	-	zh	20.00	33.13	**25.66**	53.13	29.59	37.58	31.18	**63.28**	34.98
InternLM-XComposer-2.5	InternLM2-7B	zh	23.24	**43.21**	1.22	63.59	32.62	33.92	31.55	51.72	34.44
CogVLM2	Llama3-8B	zh	27.84	39.40	11.50	62.51	**34.57**	29.26	30.56	62.81	33.37
LLaVA-Next	Llama3-8B	zh	29.73	34.36	1.33	58.14	31.98	26.58	29.85	61.88	31.00
GPT-4o	-	zh	18.11	23.87	12.39	61.60	31.25	13.91	33.57	62.19	27.95
GeoChat	Vicuna-7B	zh	24.05	22.12	0.72	28.16	23.05	21.65	17.45	38.13	20.72
LLaVA-1.5	Vicuna-7B	zh	24.05	22.12	0.66	28.05	23.01	21.65	17.45	38.12	20.70

perception and explore distillation techniques to enhance visual module performance with reduced model sizes.

Poor Performance in Visual Grounding Tasks. On XLRS-Bench, MLLMs underperform significantly, as shown in main text. On both the Chinese (XLRS-Bench-ZH) and English (XLRS-Bench-EN) benchmarks, most models achieve less than 1.0% accuracy in terms of Acc@0.5 and Acc@0.7 metrics, highlighting major limitations in their ability to handle visual localization tasks. Key issues include: 1. Inadequate local feature extraction, hindering fine-grained localization in ultra-high-resolution images. 2. Weak cross-modal alignment, limiting accurate matching between language descriptions and complex visual scenes. 3. Poor generalization to real-world remote sensing scenarios, particularly with high object similarity. 4. Limited reasoning capability (e.g., condition-based visual grounding, L-3 capability) in ultra-high-resolution settings, underscoring the need for more effective multimodal representation learning. Future research could focus on better visual feature extraction, enhanced language alignment, and stronger generalization and reasoning in complex, high-resolution contexts.

5 Sub-tasks (L-3 capability) Results on XLRS-Bench

This section highlights the performance of MLLMs across all L-3 capabilities. The VQA task is split into perception and reasoning dimensions, with results shown in Tables 2 and 3, respectively. L-3 capabilities for the Visual Grounding task are summarized in Table 4.

MLLMs generally excel in reasoning tasks compared to perception tasks in XLRS-Bench. On most benchmarks [1, 5], MLLMs excel in perception tasks but struggle with reasoning, which requires complex conditional interpretation. However, XLRS-Bench presents a reversed trend: MLLMs perform worse in perception due to its ultra-high-resolution images averaging 8,500×8,500 pixels—24 times higher than those in MME-Realworld [5]. With a 4K resolution limit, current MLLMs cannot process such detailed imagery effectively. In contrast, reasoning tasks, often based on global patterns, are less resolution-dependent. XLRS-Bench highlights the need for next-generation MLLMs capable of handling ultra-high-resolution data, a crucial step for real-world remote sensing applications.

Capturing local features is crucial for perception tasks. Ultra-high-resolution visual tasks like Object Spatial Relationship (OSR) and Object Color (OCL), as shown in Tables 2-4, exhibit sig-

Table 3: **Experimental results of L-3 capability on the reasoning dimension of VQA tasks.** Models are ranked according to their average performance. Rows corresponding to proprietary models are highlighted in gray for distinction. “RP”, “AD”, “ECR”, “CCR” and “RCCD” each indicate a specific task domain: Route Planning, Anomaly Detection, Environmental Conditional Reasoning, Counting with Complex Reasoning and Regional Counting with Change Detection.

Method	**LLM**	**Language**	**Reansoning**					
Subtasks (L-3 Capability)			**RP**	**AD**	**ECR**	**CCR**	**RCCD**	**Avg**
InternVL2	InternLM2.5-7B	en	33.01	**74.54**	77.07	**54.12**	44.44	**58.97**
InternLM-XComposer-2.5	InternLM2-7B	en	35.31	69.50	77.87	49.59	32.22	56.83
Qwen2-VL	Qwen2-7B	en	32.12	68.35	79.29	46.81	**45.93**	56.33
LLaVA-Next	Llama3-8B	en	26.02	69.10	76.00	45.47	32.22	53.14
LLaVA-OneVision	Qwen2-7B	en	24.07	71.88	**79.91**	37.76	37.78	53.00
GPT-4o	-	en	**41.24**	72.06	75.29	26.34	21.85	52.79
CogVLM2	Llama3-8B	en	34.16	69.85	73.07	45.37	-	52.70
GPT-4o-mini	-	en	33.81	72.06	75.02	31.49	15.19	51.60
LLaVA-1.5	Vicuna-7B	en	38.67	34.70	41.51	25.00	29.26	34.97
GeoChat	Vicuna-7B	en	32.12	33.25	34.84	25.10	-	29.71
InternVL2	InternLM2.5-7B	zh	40.09	**76.57**	**85.51**	48.97	44.07	**62.14**
Qwen2-VL	Qwen2-7B	zh	24.34	**76.57**	83.11	**49.69**	**44.44**	57.89
LLaVA-OneVision	Qwen2-7B	zh	29.12	74.80	82.13	39.92	30.37	55.51
GPT-4o-mini	-	zh	**41.86**	74.71	73.78	34.05	21.85	54.84
InternLM-XComposer-2.5	InternLM2-7B	zh	29.47	69.14	80.89	42.28	24.44	54.06
CogVLM2	Llama3-8B	zh	26.19	70.47	75.47	41.15	-	50.60
LLaVA-Next	Llama3-8B	zh	21.59	69.85	73.51	31.58	25.56	48.33
GPT-4o	-	zh	27.08	69.41	66.76	20.78	15.19	45.05
LLaVA-1.5	Vicuna-7B	zh	38.67	37.58	41.60	25.10	29.26	35.72
GeoChat	Vicuna-7B	zh	22.79	23.74	24.44	25.10	-	22.58

Table 4: Visual grounding performance of L-3 capability on XLRS-Bench.

L-3 Capability	**Language**	**Method**	**GPT-4o**	**GPT-4o-mini**	**Qwen2-VL**	**LLaVA-OneVision**	**LLaVA-Next**	**LLaVA-1.5**	**CogVLM2**	**InternLM-XComposer-2.5**	**InternVL2**	**GeoChat**
Fine-grained Visual Grounding	en	Acc@0.5	**0.70**	0.17	0.21	0.25	0.16	0.11	0.02	0.03	0.46	0.21
		Acc@0.7	0.10	0.06	0.03	0.00	0.08	0.00	0.00	0.02	**0.17**	0.02
Condition-based Visual Grounding	en	Acc@0.5	**0.21**	0.00	0.08	0.06	0.19	0.06	0.00	0.00	0.19	0.06
		Acc@0.7	0.00	0.00	0.02	0.00	0.00	0.00	0.00	0.00	**0.06**	0.00
Fine-grained Visual Grounding	zh	Acc@0.5	**0.76**	0.22	0.22	0.24	0.05	0.13	0.02	0.08	0.38	0.21
		Acc@0.7	0.05	0.05	0.02	0.02	0.00	0.02	0.00	0.00	**0.11**	0.02
Condition-based Visual Grounding	zh	Acc@0.5	0.14	**0.20**	0.06	0.02	0.08	0.11	0.03	0.03	0.17	0.06
		Acc@0.7	0.00	0.00	0.00	0.00	**0.03**	0.02	0.00	0.00	**0.03**	0.00

nificantly lower accuracy. For example, LLaVA1.5 achieves only 17.45% on the OCL task in the Chinese benchmark. This shortfall arises from three main factors: 1. Limited sensitivity to sparse details. Ultra-high-resolution images (e.g., 8,500×8,500 pixels) feature sparse yet critical details, such as small object contours and intricate local relationships. Existing MLLMs, optimized for global feature extraction, struggle to detect these fine-grained signals, impairing performance on detail-oriented tasks. 2. Imbalanced global and local feature modeling. Current visual encoders emphasize global semantics over local-global interactions. For instance, RC tasks require counting objects in small regions, but an overreliance on global features often results in missed or miscounted local targets. 3. Resolution and computational constraints. MLLMs face input resolution limits (e.g., 4K), necessitating downsampling that degrades local features and hampers detailed modeling. Ultra-high-resolution scenarios demand enhanced local feature modeling to meet their complex requirements. Strengthening this capability is essential for improving performance in these perception tasks.

6 Samples and Hard Cases of XLRS-Bench

In this section, we present examples of the VQA (Fig. 2), image captioning (Fig. 4 and Fig. 5), and visual grounding tasks (Fig. 3). What's more, we construct a detailed table (Tab. 5) analyzing model performance and error causes for each L-3 subtask. We then use examples to thoroughly illustrate the errors for each subtask.

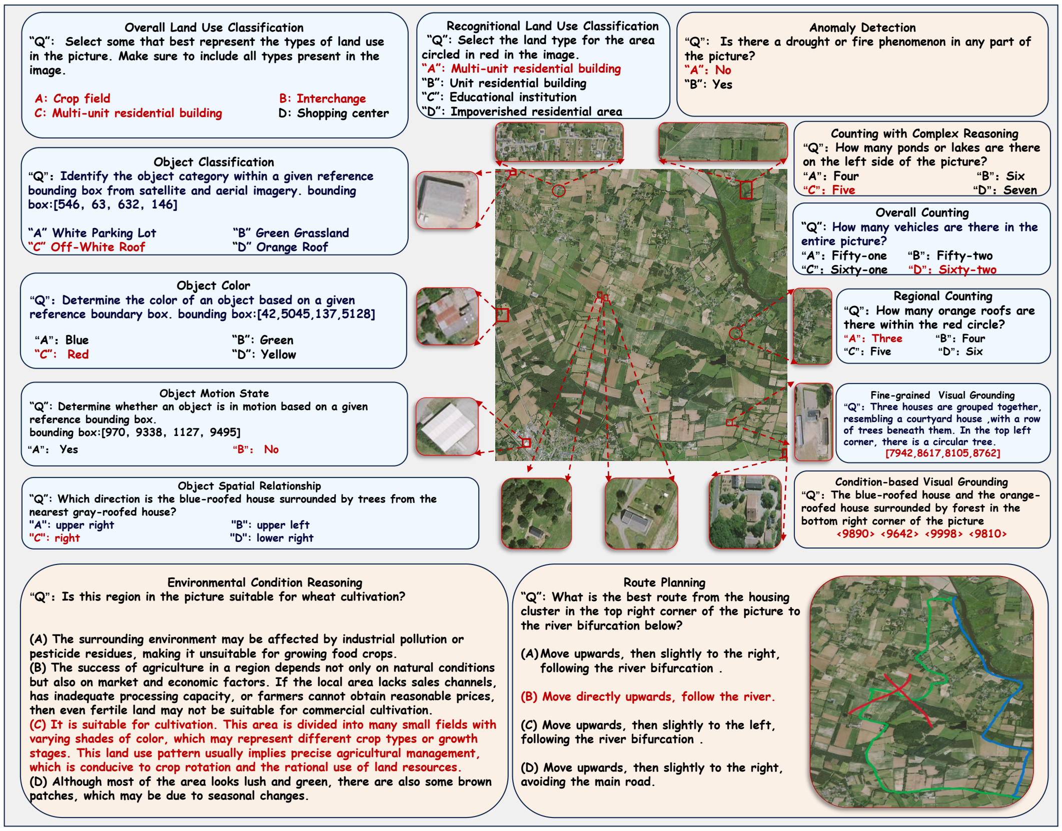
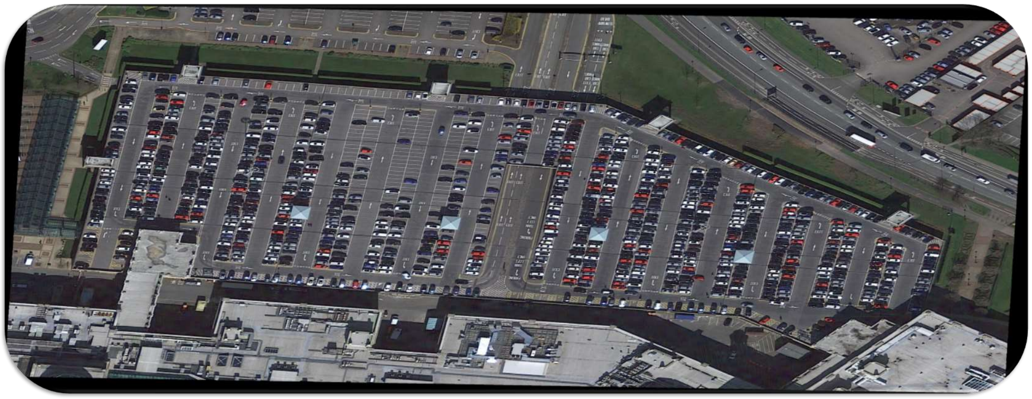


Figure 2: Example of XLRS-Bench in English. XLRS-Bench focuses on large-size ultra-high-resolution remote sensing imagery, integrating over 10 multimodal perception and reasoning tasks within the same image.

Figure 3: **Visual Grounding Results of XLRS-Bench.** Question:*"The multi-sided building in the left central area of the picture."* The "multi-sided building" required in the ground truth is relatively small and therefore difficult to identify. The GPT-4o model incorrectly classified a similarly shaped roundabout as a polygonal building, while GeoChat [6] misidentified an irregularly shaped parking lot as a polygonal building.



Ground truth

This remote sensing image shows a complex urban area, including parking lots, roads, buildings, and green areas. The image is divided into nine parts from left to right and from top to bottom, each with its unique features and elements.

1. In the top left corner, a parking lot is shown with multiple cars, featuring a variety of colors, mainly standard sedans and SUVs. To the north of the parking lot is a narrow road with a white van driving on it. The left edge has a green grassy area, with no apparent buildings or other structures.
2. The upper middle part displays another parking lot with multiple cars, mainly in white, black, or gray colors. There is a small green vegetation area at the south end of the parking lot. To the right is a wide road with clear lane division lines and direction signs.
3. The top right corner depicts a busy traffic scene, including a two-way four-lane road and a parking lot. There are multiple cars on the road and in the parking lot, with a variety of car colors.
4. The middle-left part shows a large parking lot with multiple cars, predominantly in dark colors. The left edge of the parking lot has a row of green trees, and there is a construction area in the lower right corner.
5. The middle part displays a large parking lot with multiple cars, with a wide two-way road in the middle of the parking lot marked with direction arrows and the word "EXIT".
6. The middle right part shows a parking lot and a traffic road. There are multiple cars in the parking lot, and the road is a two-way four-lane highway.
7. The bottom left corner displays a complex urban area, including a large building and surrounding parking lots. The building's roof is covered with mechanical equipment, and there are multiple cars parked around.
8. The lower middle part shows a commercial or industrial area, with a large building and multiple neatly parked vehicles. There are narrow passages between the buildings.
9. The bottom right corner displays a parking lot and two buildings. There are multiple cars in the parking lot, and the buildings have gray-white roofs with irregular shapes.

Overall, this image shows a highly developed urban area with a high density of parked vehicles and building layouts. The road shapes are generally straight, with moderate widths, and the building roofs are mostly gray-white. From the image, it can be inferred that the area might be a commercial or industrial center, with high traffic flow and a complex transportation system.

LLaVA-OneVision

The image is an aerial view of a large parking lot adjacent to a highway. The parking lot is filled with numerous cars, and there are several buildings nearby. The highway has multiple lanes with vehicles moving along it. There are also some green spaces visible in the vicinity.

1. 1. The top left section shows a few cars parked in the parking lot, with a building partially visible on the left side.
2. The top middle section displays a row of cars parked in the parking lot, with a road running parallel to the parking area.
3. The top right section shows a few cars parked in the parking lot, with a building visible on the right side. (Lack of number of lanes)
4. The middle-left section shows a large open space in the parking lot, with a few cars parked around it.
5. The middle section shows a row of cars parked in the parking lot, with a building visible on the left side. (Lack of description of details in the image content)
6. The middle right section shows a few cars parked in the parking lot, with a building visible on the right side. (Lack of number of lanes)
7. The bottom left section shows a large open space in the parking lot, with a few cars parked around it.
8. The bottom middle section shows a row of cars parked in the parking lot, with a building visible on the left side
9. The bottom right section shows a few cars parked in the parking lot, with a building visible on the right side.

Road Description: The highway runs parallel to the parking lot and has multiple lanes with vehicles moving along it. There are also some roads running perpendicular to the highway, connecting to the parking lot.

Building Characteristics: There are several buildings visible in the image, including a large building adjacent to the parking lot and smaller buildings scattered around the area. The buildings have different roof colors and sizes.

Airport Details: There are no visible signs of an airport in the image.

Comprehensive Inference: The image depicts a large parking lot adjacent to a highway, with numerous cars parked in it. There are several buildings nearby, and the highway has multiple lanes with vehicles moving along it. The overall functional characteristics of the image suggest that it could be a commercial or industrial area, given the presence of the large building and the number of cars parked in the lot. The layout of the major roads suggests good connectivity to other areas. There are no visible signs of human activity other than the parked cars. (Lack of reasoning to summarize the content of the entire image)

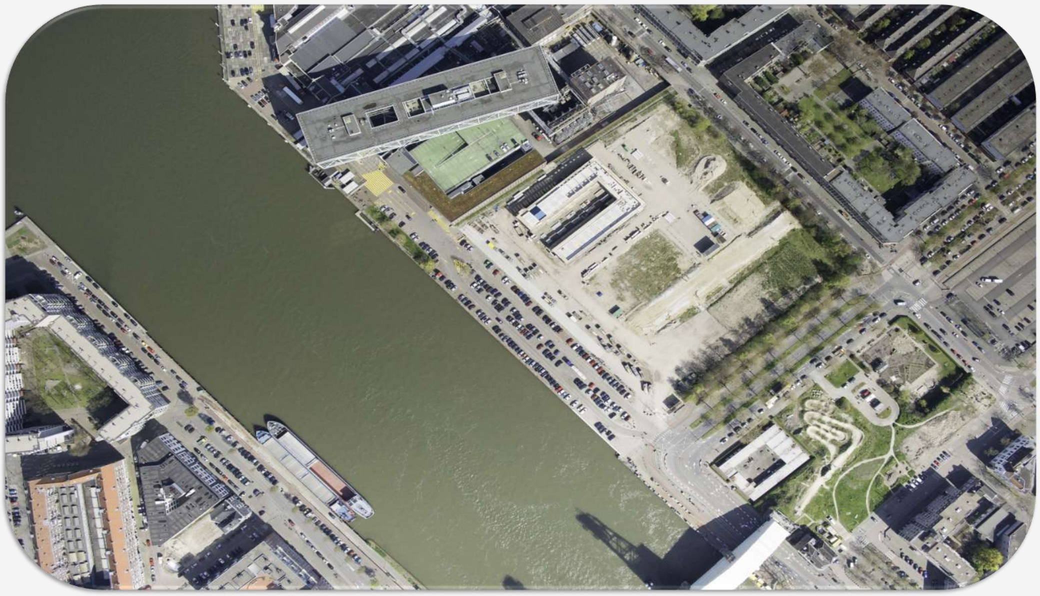
LLaVA-Next

The image is a remote sensing photograph of a large parking lot situated near a highway. The parking lot is filled with numerous cars parked in rows, and there are several buildings nearby. The highway runs parallel to the parking lot, and there are trees and green spaces scattered throughout the area.

1. The leftmost section of the image shows a portion of the parking lot with several cars parked in rows. There are no distinct buildings or other objects in this section.
2. The second section from the left shows a larger portion of the parking lot with more cars parked. There are a few trees visible in this section.
3. The third section from the left shows a smaller portion of the parking lot with fewer cars. There are no distinct buildings or other objects in this section. (Lack of number of lanes)
4. The fourth section from the left shows a larger portion of the parking lot with more cars parked. There are a few trees visible in this section.
5. The fifth section from the left shows a smaller portion of the parking lot with fewer cars. There are no distinct buildings or other objects in this section. (Lack of description of details in the image content)
6. The sixth section from the left shows a larger portion of the parking lot with more cars parked. There are a few trees visible in this section. (Lack of number of lanes)
7. The seventh section from the left shows a smaller portion of the parking lot with fewer cars. There are no distinct buildings or other objects in this section.
8. The eighth section from the left shows a larger portion of the parking lot with more cars parked. There are a few trees visible in this section.
9. The ninth section from the left shows a smaller portion of the parking lot with fewer cars. There are no distinct buildings or other objects in this section

The image shows a large parking lot with numerous cars parked in rows, surrounded by a few trees and green spaces. There are several buildings nearby, and a highway runs parallel to the parking lot. The image does not show any distinct industrial or residential areas, but the presence of the parking lot and highway suggests that this area may be a commercial or transportation hub. The overall layout of the image suggests a well-organized and functional transportation network, with the parking lot and highway serving as key components. The image does not provide any clear signs of human activity or ecological elements beyond the parking lot and highway. There are no visible bodies of water or other natural features in the image. The image does not provide any information about the number of vehicles or ships present in the area. (Lack of reasoning to summarize the content of the entire image)

Figure 4: **Image Captioning Results of XLRS-Bench in English.** LLaVA-OneVision and LLaVA-Next face challenges in capturing image details, particularly in conveying critical information like lane counts and vehicle types. Their descriptions often lack depth, failing to convey the richness and nuances of the images. The language is overly rigid and mechanical, struggling to naturally align with the images' context and overall environment.



Ground truth

这张遥感图像展示了一个城市区域，包含河流、建筑群、道路、停车场和绿地。图像从左至右、从上至下被分为九个部分，每个部分展示了城市的不同功能和特征。
1.在图像的左上部分，显示了一条河流和河岸边的停车场及建筑物。河流呈深灰绿色，沿着图像的中部水平延伸。停车场内停有约40辆车，颜色包括白色、黑色、灰色和红色。建筑物的屋顶为灰白色，形状为矩形。
2.中上部分展示了一个带有太阳能板的长条形建筑，下方停有多辆主要为黄色和白色的车辆，右侧开放停车场中车辆主要为黑色、白色和灰色。
3.右上部分是一条河流旁的道路，上有多辆车辆行驶，道路呈直线形状，宽度均匀。河中有一艘大型游船，船身为白色。
4.左中部分显示了一个城市交叉口，包括多条道路、车辆和树木。交叉口中有车辆，颜色包括蓝色、白色、黑色和红色。
5.中间部分是一个复杂的建筑群，包括居住和休闲设施，周围环境由道路和少量绿地组成。
6.右中部分展示了一个城市河岸景观，包括河流、停车场、直道路和建筑群。停车场内有约100辆车，建筑物屋顶为灰色和白色。
7.左下部分描述了一个靠近水体的建筑区域，包括建筑、停车场和道路。停车场内约有几十辆车，车辆颜色多样。
8.中下部分是一个城市区域，包括建筑物、道路、绿地和停车场。左下角和右下角的停车场内分别停有15辆和12辆车。
9.右下部分展示了一条宽阔的河流和沿河道路，道路上有约十辆车行驶，车辆颜色多样。
通过这些描述，可以推理出图像展示了一个城市的多功能区域，包括商业、居住和休闲设施。河流和道路的存在表明该城市可能依水而建，且交通发达。**建筑物的密集程度和停车场的充足表明这是一个人口密集且经济活跃的区域。**

LLaVA-OneVision

图像展示了一个城市河岸区域，包括建筑物、道路、绿地和水体。主要元素包括多层建筑、停车场、河流和一些绿色植被。分区描述：
1. 左上角：多层建筑，屋顶为灰色，周围有停车场，车辆颜色多样。（Lack of regional counting）
2. 中上部分：河流沿岸，河岸上有绿色植被和一些小型建筑。
3. 右上部分：道路平行于河流，车辆沿着道路行驶。
4. 左中部分：河流对岸，有更多建筑和停车场。
5. 中间部分：河流中央有一艘船只，河岸两侧有绿色植被。（Lack of regional counting）
6. 右中部分：道路与河流平行，车辆沿着道路行驶。
7. 左下部分：多层建筑，屋顶为红色，周围有停车场。
8. 中下部分：河流沿岸，河岸上有绿色植被和一些小型建筑。（Lack of regional counting）
9. 右下部分：道路与河流平行，车辆沿着道路行驶。（Lack of regional counting）

综合推理：－ 功能分区描述：图像显示了工业区（左侧多层建筑）、住宅区（右侧多层建筑）和绿地（河流沿岸的绿色植被）。－ 交通网络描述：主要道路平行于河流，车辆沿着道路行驶，形成了城市的交通网络。－ 生态与人类活动：河流和绿地提供了生态功能，而多层建筑和停车场则表明这是一个居住和商业混合的区域。河流中央的船只可能用于运输或休闲活动。－ 常见物体计数：在图像中可以看到数十辆汽车沿着道路行驶，以及河岸边的几艘船只。（Lack of reasoning to summarize the content of the entire image）

LLaVA-Next

这张遥感图像展示了一个城市的部分区域，主要包括水域、道路、建筑物、绿地和一些交通工具。图像的视角是从高空拍摄的，提供了一个鸟瞰的视野。
1. 从左到右，第一部分显示了一个水域，水面平静，水域的颜色呈现出一种绿色。（Lack of regional counting）
2. 第二部分显示了一个道路，道路呈现出一种灰色，道路的形状是直线，宽度较宽，长度较长，方向是从左到右。
3. 第三部分显示了一个建筑物，建筑物呈现出一种灰色，建筑物的形状是矩形，大小较大，位置在水域的右侧。
4. 第四部分显示了一个绿地，绿地呈现出一种绿色，绿地的形状是长方形，大小较大，位置在建筑物的右侧。
5. 第五部分显示了一个建筑物，建筑物呈现出一种灰色，建筑物的形状是矩形，大小较大，位置在绿地的右侧。（Lack of regional counting）
6. 第六部分显示了一个道路，道路呈现出一种灰色，道路的形状是曲线，宽度较宽，长度较长，方向是从左到右。
7. 第七部分显示了一个建筑物，建筑物呈现出一种灰色，建筑物的形状是矩形，大小较大，位置在道路的右侧。
8. 第八部分显示了一个绿地，绿地呈现出一种绿色，绿地的形状是长方形，大小较大，位置在建筑物的右侧。（Lack of regional counting）
9. 第九部分显示了一个建筑物，建筑物呈现出一种灰色，建筑物的形状是矩形，大小较大，位置在绿地的右侧。（Lack of regional counting）
整体上，这张遥感图像展示了一个城市的部分区域，主要包括水域、道路、建筑物、绿地和一些交通工具。交通工具包括了一个船只和一些车辆。道路的布局呈现出一种复杂的交通网络，包括直线和曲线的道路。建筑物的密度较高，包括了各种功能区类型。绿地的分布较均匀，包括了草地和树木。水域的颜色呈现出一种绿色，表明水域的水质较好。这张图像提供了一 个城市的鸟瞰视野，展示了城市的交通、建筑和生态等方面的信息。（Lack of reasoning to summarize the content of the entire image）

Figure 5: **Image Captioning Results of XLRS-Bench in Chinese.** Both LLaVA-Next and LLaVA-OneVision share a key weakness: the inability to perform local counting. LLaVA-Next shows notable limitations and a narrow focus when describing localized features, resulting in overly simplistic outputs. While LLaVA-OneVision offers greater diversity and detail, it still lacks sufficient complexity. More importantly, both models are confined to directly describing objects in images without engaging in deeper reasoning or analysis, restricting their practical utility.

Table 5: Table index of case study figures by sub-tasks (L-3 capability) with associated (error) categories for each MLLM.

Case Figure	L-2 task	L-3 task	LLaVa-Next	Qwen2-VL	LLaVA-OneVision
Fig. 6	Anomaly detection	Anomaly detection and interpretation	**Lack of Capability**	**Reasoning Error**	**Lack of Capability**
Fig. 7	Complex reasoning	Environmental condition reasoning	**Reasoning Error**	**Correct**	**Correct**
Fig. 8	Planning	Route planning	**Lack of Capability**	**Lack of Capability**	**Lack of Capability**
Fig. 9	Spatiotemporal reasoning	Counting with change detection	**Lack of Capability**	**Fail to Follow Instruct**	**Lack of Capability**
Fig. 10	Complex reasoning	Counting with complex reasoning	**Lack of Capability**	**Fail to Follow Instruct**	**Correct**
Fig. 11	Counting	Overall counting	**Lack of Capability**	**Lack of Capability**	**Lack of Capability**
Fig. 12	Counting	Regional counting	**Lack of Capability**	**Lack of Capability**	**Lack of Capability**
Fig. 13	Scene classification	Overall land use classification	**Perception Error**	**Correct**	**Perception Error**
Fig. 14	Scene classification	Regional land use classification	**Perception Error**	**Correct**	**Correct**
Fig. 15	Object properties	Object classification	**Correct**	**Correct**	**Lack of Capability**
Fig. 16	Object properties	Object color	**Correct**	**Perception Error**	**Perception Error**
Fig. 17	Object properties	Object motion state	**Perception Error**	**Lack of Knowledge**	**Lack of Knowledge**
Fig. 18	Object spatial relationship	Object spatial relationship	**Perception Error**	**Perception Error**	**Correct**

In this section, we present a case study analysis of the error types made by LLaVa-Next, Qwen2-VL, and LLaVA-OneVision on various sub-tasks in XLRS-Bench. We classify the errors into the following 5 categories, following the MMT-Bench [7]:

Perception Error : MLLMs often struggle to recognize, classify, or detect objects and content in images, largely due to the limited representational power of their visual encoders, making this the most prevalent error. This perceptual limitation is especially evident in ultra-high-resolution images, where MLLMs often struggle to detect objects with minimal pixel representation. See examples in Fig. 17, Fig. 18, etc.

Reasoning Error : MLLMs accurately perceive visual content but fail in reasoning, resulting in incorrect answers.. See examples in Fig. 6, Fig. 7, etc.

Lack of Knowledge : MLLMs lack the domain-specific knowledge needed to answer specialized questions, such as identifying ship wake information in remote sensing images (see Fig. 17).

Lack of Capability : MLLMs do not have the capability to solve the corresponding tasks. See examples in Fig. 6, Fig. 8.

Fail to Follow Instruct : MLLMs often misinterpret instructions, leading to errors. For instance, they may misunderstand specific conditions (see Fig. 9) or disregard the instructions entirely, generating errors for the image instead (see Fig. 11).

Anomaly detection and interpretation

Question: Why are there many industrial areas clustered together on the left side of the highway in the lower left of the picture?

Options
(A) Because the region is close to railways or airports, it facilitates the transportation of raw materials and finished products.',
(B) The region has abundant natural resources, such as coal and iron, suitable for the development of heavy industry.',
(C) This layout may be the result of urban planning, aimed at separating industrial activities from agricultural production areas to reduce environmental impact.',
(D) The regional government has offered preferential policies, attracting a large number of enterprises to invest in building factories.

Answer: C

10000×10000

Lack of Capability

LLaVA-Next:A

The correct answer is (A) Because the region is close to railways or airports, it facilitates the transportation of raw materials and finished products. The image shows a landscape with a highway running through it, and on the left side of the highway, there are many industrial areas clustered together. This layout is likely due to the presence of transportation infrastructure such as railways or airports nearby. These transportation networks make it easier and more cost-effective to move raw materials and finished products, which is essential for the operation of industrial facilities. The concentration of industrial areas in this region suggests that the transportation infrastructure plays a significant role in the development and growth of the industrial sector in this area.

Reasoning Error

Qwen2-VL:D

The correct answer is (D) The regional government has offered preferential policies, attracting a large number of enterprises to invest in building factories. The image shows a mix of agricultural fields and industrial areas. The clustering of industrial areas on the left side of the highway in the lower left of the picture suggests a deliberate layout. This type of arrangement is often seen in regions where the government has implemented policies to attract industries. Preferential policies can include tax incentives, land allocation, and infrastructure development, which make the area more attractive for businesses. This strategic placement helps in the economic development of the region by creating jobs and promoting industrial growth.

Lack of Capability

LLaVA-OneVision: A

The clustering of industrial areas near the highway suggests a strategic location that facilitates transportation. Proximity to major transport routes like highways often enhances the efficiency of moving raw materials and finished products, making option (A) the most plausible explanation.

Figure 6: A sample case of Anomaly detection and interpretation.

Environmental condition reasoning

Question: According to satellite images, the area covered by greenery and land is extensive and evenly distributed, yet the residential areas are very dense. Why does this phenomenon occur?

Options

(A) This layout may be due to the fact that the region has historically been an agricultural area, thus preserving a large amount of farmland and green spaces. Over time, as urbanization has accelerated, the original agricultural land has been retained, forming the current landscape.

(B) The planners of the region may believe that dispersed residential areas and concentrated green spaces can create a harmonious living environment, bringing people closer to nature and allowing them to enjoy a tranquil lifestyle.

(C) Concentrated residential areas are likely located near transportation hubs or business centers, facilitating residents' travel and work, while also providing the necessary space for urban development.

(D) This layout may be intended to protect the natural environment and ecological balance, while also meeting human residential needs. Large areas of green space can provide more ecosystem services, such as air purification and water conservation, while concentrated residential development can enhance land use efficiency and reduce the impact on the surrounding environment.

Answer: D

10000×10000

Reasoning Error

LLaVA-Next: A

The correct answer is (A) This layout may be due to the fact that the region has historically been an agricultural area, thus preserving a large amount of farmland and green spaces. Over time, as urbanization has accelerated, the original agricultural land has been retained, forming the current landscape.

Qwen2-VL: D

The correct answer is (D).The image shows a mix of agricultural land, green spaces, and residential areas. The extensive greenery and land patches suggest that the region has historically been agricultural, which has been preserved over time. The dense residential areas indicate that urbanization has occurred, but the green spaces have been maintained, forming a harmonious landscape.

LLaVA-OneVision: D

The image shows a landscape where there are extensive areas of greenery and land, but also densely populated residential areas. This phenomenon can be attributed to several factors: Historical Context: The region may have historically been an agricultural area, with large tracts of land dedicated to farming. Over time, as urbanization accelerated, the original agricultural land was retained, forming the current landscape. This is evident in the large patches of green fields that are interspersed with residential areas.

Figure 7: A sample case of Environmental condition reasoning.

Route planning

Question: What is the shortest route from the triangular road above the largest cluster of houses in the top right corner of the picture to the nearest roundabout below it to the left?

Options

(A) Starting from the intersection at the upper left of the triangular road, head south along the road, turn right at the fourth intersection, continue west, turn left at the third intersection, proceed to the end of the road, turn left, and continue for a while to reach your destination.',

(B) Starting from the intersection at the upper left of the triangular road, head south along the road, turn right at the fourth intersection, continue west, turn left at the third intersection, proceed to the end of the road, turn right, and continue straight for a while to reach your destination.',

(C) Starting from the intersection at the upper left of the triangular road, head south along the road, turn left at the first intersection, continue along the road, turn left at the next intersection, and proceed for a while to reach your destination.',

(D) Starting from the intersection at the upper left of the triangular road, head south along the road, pass four intersections, and continue for a while to reach your destination.

Answer: D

10000×10000

Lack of Capability

LLaVA-Next:A

The shortest route from the triangular road above the largest cluster of houses in the top right corner of the picture to the nearest roundabout below it to the left is option (A). Starting from the intersection at the upper left of the triangular road, you head south along the road. At the fourth intersection, you turn right. Continuing west, you turn left at the third intersection. Proceeding to the end of the road, you turn left, and continue for a while to reach your destination. This route avoids unnecessary turns and takes the most direct path to the roundabout.

Lack of Capability

Qwen2-VL:B

The shortest route from the triangular road above the largest cluster of houses in the top right corner of the picture to the nearest roundabout below it to the left is (B). Starting from the intersection at the upper left of the triangular road, head south along the road, turn right at the fourth intersection, continue west, turn left at the third intersection, proceed to the end of the road, turn right, and continue straight for a while to reach your destination.

Lack of Capability

LLaVA-OneVision: B

The shortest route from the triangular road above the largest cluster of houses in the top right corner of the picture to the nearest roundabout below it to the left would be option B. This route involves traveling south along the road, passing four intersections, and continuing for a while until reaching the roundabout.

Figure 8: A sample case of Route planning.

Counting with changing detection

10000×10000

Question: How many changes are there in the number of houses within the red-circled areas corresponding to the two pictures?

Options
(A) 4
(B) 3
(C) 6
(D) 5

Answer: D

Lack of Capability

LLaVA-Next:B

The correct answer is (B) .To explain, let's analyze the red-circled areas in the two images: 1. In the top image, there are three distinct areas with houses within the red circles. 2. In the bottom image, there are also three distinct areas with houses within the red circles. The number of houses within the red-circled areas remains consistent across the two images, which is why the correct answer is (B) 3.

Fail to Follow Instruct

Qwen2-VL:B

The correct answer is (B) 3. To determine the number of changes in the number of houses within the red-circled areas corresponding to the two pictures, we need to compare the houses in each area in both images.\n\n1. **First Image:**\n - Red-circled area 1: 2 houses\n - Red-circled area 2: 3 houses\n - Red-circled area 3: 1 house\n - Red-circled area 4: 2 houses\n\n2. **Second Image:**\n - Red-circled area 1: 3 houses\n - Red-circled area 2: 4 houses\n - Red-circled area 3: 2 houses\n - Red-circled area 4: 3 houses

Lack of Capability

LLaVA-OneVision: B

The answer is B.

Figure 9: A sample case of Counting with change detection.

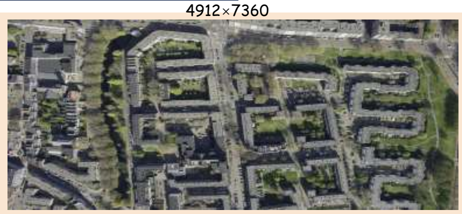
Counting with complex reasoning

Question: How many cars are on the road in the downward direction from the intersection in the upper right corner of the picture??

Options
(A) 7
(B) 6
(C) 4
(D) 5

Answer: B

4912×7360



Lack of Capability

LLaVA-Next:C

The correct answer is (C) . To count the number of cars on the road in the downward direction from the intersection in the upper right corner of the picture, we need to look at the road that appears to be a one-way street. From the intersection, we can see that the road is clear and there are four cars visible on this road. The cars are aligned in a straight line, indicating that they are moving in the same direction. The cars are not overlapping, so we can count them accurately. The other roads in the image are not in the downward direction from the intersection, so they do not contribute to the count.

Lack of Capability

Qwen2-VL:C

The correct answer is (C) . In the image, there is an intersection in the upper right corner. From this intersection, the road goes downward in a straight line. Counting the cars on this road, there are exactly 4 cars visible.

Lack of Capability

LLaVA-OneVision: A

There are 7 cars on the road in the downward direction from the intersection in the upper right corner of the picture.

Figure 10: A sample case of Counting with complex reasoning.

Overall counting

Question: How many cars are there in the whole picture?

Options
(A) 107
(B) 172
(C) 144
(D) 165

Answer: D

6000×6000

Lack of Capability

LLaVA-Next:C

The correct answer is (C) 144. The image shows a large parking lot with multiple rows of cars parked in an orderly manner. To count the number of cars, I would need to know the exact dimensions of the parking lot and the number of cars in each row. However, without that information, I can only provide an estimate based on the visible area. From the image, it appears that there are at least 144 cars visible. The actual number could be higher if there are additional rows of cars not visible in the image or if the cars are packed more tightly than they appear in the photo.

Fail to Follow Instruct

Qwen2-VL:B

The correct answer is (B) 172. To determine the number of cars in the picture, we need to count the cars in each visible parking area and on the roads. Here is the breakdown:\n\n1. **Parking Areas:**\n - Top right corner: 10 cars\n - Top left corner: 5 cars\n - Bottom left corner: 10 cars\n - Bottom right corner: 15 cars\n - Middle left area: 10 cars\n - Middle right area: 10 cars\n - Middle bottom area: 10 cars\n - Middle top area: 10 cars\n - Middle bottom right area: 10 cars\n - Middle bottom left area: 10 cars\n - Middle top left area: 10 cars\n - Middle top right area: 10 cars\n - Middle bottom center area: 10 cars\n - Middle top center area: 10 cars\n - Middle bottom center right area: 10 cars\n

LLaVA-OneVision: D

There are 165 cars in the picture.

Figure 11: A sample case of Overall counting.

Regional counting

Question: How many cars are there inside the red circle?

Options
(A) 17
(B) 178
(C) 19
(D) 20

Answer: D

4912×7360

Lack of Capability

LLaVA-Next:C

The correct answer is (C) .To explain, the image shows an aerial view of a city with various buildings, roads, and parking areas. The red circle is drawn around a specific area within the city. Inside this circle, there are 19 cars visible. The cars are parked along the sides of the roads and in designated parking spaces. The number of cars can be counted by visually identifying each car within the red circle. The other options (A, B, and D) are incorrect because they do not match the number of cars visible in the red circle.

Lack of Capability

Qwen2-VL:A

The correct answer is (A) . To determine the number of cars inside the red circle, we need to carefully count each car within the designated area. The red circle is drawn around a specific section of the image, and by examining the area within this circle, we can see a total of 17 cars.

Lack of Capability

LLaVA-OneVision: A

There are 17 cars inside the red circle.

Figure 12: A sample case of Regional counting.

Overall Land use classification

4096×4096

Question: Choose some land use types that best describes the image. Requires all types contained in the picture to be selected.

Options
(A) Port
(B) Aquaculture
(C) Burial Site
(D) Crop Field

Answer: ABD

Perception Error

LLaVA-Next:ABCD

The image shows a coastal area with various land use types. The correct answer is:(A) Port (B) Aquaculture (C) Burial Site (D) Crop Field. The image shows a coastal area with a body of water, which appears to be a river or a bay. There are several boats visible on the water, indicating that this is a port or a harbor. The presence of a large building with a dome-like structure suggests that there might be a significant industrial or commercial activity in the area. The land is divided into different sections, some of which are likely used for agricultural purposes, as indicated by the rectangular fields. There are also areas that seem to be used for residential purposes, as evidenced by the clusters of buildings. The presence of a road network suggests that this is a developed area with infrastructure. The overall layout and the presence of a large body of water suggest that this is a port or a harbor area. Aquaculture is also a possibility, as there are areas that could be used for fish farming. However, without more specific details, it is not possible to confirm the presence of a burial site. Crop fields are also present, as indicated by the rectangular fields.

Qwen2-VL:ABD

Correct answer: (A) Port, (B) Aquaculture, (D) Crop Field

Perception Error

LLaVA-OneVision: ABCD

The image shows a river with boats, indicating that it is navigable and possibly used for transportation or recreation. There are industrial buildings along the riverbank, suggesting that this area may be used for manufacturing or other industrial activities. The presence of green areas indicates that there are parks or undeveloped land in the vicinity. The grid-like pattern of the land suggests that some parts of the area are used for agriculture, as these are often organized in such patterns for efficient farming practices. Therefore, the correct options are A (Port), B (Aquaculture), D (Crop Field), and potentially C (Burial Site) if there are any cemeteries in the area.

Figure 13: A sample case of Overall land use classification.

Regional Land use classification

10000×10000

Question: Select the land use types in the areas circled in red within the image.

Options
(A) Golf Course
(B) Park
(C) Recreational Facilities
(D) Stadium

Answer: D

Perception Error

llama3-llava-next-8b: A

The circled areas in the image appear to be well-maintained and have a pattern that is typical of a golf course. Golf courses are designed with a series of holes, each consisting of a tee box, fairway, rough, and putting green, with a flagstick and a hole. The layout and the green spaces suggest that these areas are likely designated for golfing activities. The presence of a body of water nearby also indicates that the golf course might have water hazards, which are common features in golf course designs.

Qwen2-VL: C

The correct answer is (C) . The areas circled in red appear to be small, rectangular plots of land with a more uniform and organized appearance compared to the surrounding agricultural fields. This suggests that they are likely to be recreational facilities, such as sports fields, parks, or other types of recreational areas. The uniformity and the presence of what appears to be a small building or structure within these areas further support this classification.

LLaVA-OneVision: C

The land use type in the area circled in red appears to be a recreational facility. This is indicated by the presence of open spaces, pathways, and possibly sports fields or other recreational amenities. The layout suggests it is designed for public use, which aligns with the characteristics of a recreational facility.

Figure 14: A sample case of Regional land use classification.

Object classification

Question: Recognize the category of objects from satellite and aerial images given the bounding boxes for referring objects.
The resolution of satellite image is 6000 x 6000. Bounding box: [2079,3574,2238,3637]

Options
(A) white van
(B) blue vehicle
(C) red truck
(D) grey buildings

Answer: D

6000×6000

LLaVA-Next: A

The correct answer is (A) white van. The bounding box [2079,3574,2238,3637] is likely a white van due to its size and color. The resolution of the satellite image is 6000 x 6000, which allows for a clear distinction between different types of vehicles. The white van is distinguishable from other vehicles such as the blue vehicle, red truck, and grey buildings. The van's color and shape are more prominent in the image, making it easier to identify. The other bounding boxes are for different types of objects, such as buildings and vehicles, but the one in question is specifically for a white van.

Qwen2-VL: A

The correct answer is (A) white van. The image shows an aerial view of a city area with various buildings, roads, and vehicles. The bounding box [2079,3574,2238,3637] is located on the ground, and within this area, there is a white van parked on the street. The van is clearly visible and identifiable by its white color and shape, making it the correct choice.

Lack of Capability

LLaVA-OneVision: D

The bounding box you provided [2079,3574,2238,3637] corresponds to a small section of the image. Based on the context and the objects visible in that specific area, it appears to be part of a road or pathway. The resolution of 6000 x 6000 suggests that this is a zoomed-in view of a larger area, which includes residential buildings with red roofs, as seen in the surrounding areas. The objects within the bounding box do not seem to match any of the options A, B, C, or D directly. However, since the bounding box is so small and only shows a portion of the image, it\'s challenging to determine the exact nature of the object without more context. If we were to interpret the bounding box as a single object, it might be considered a "grey building" (D) due to its proximity to the buildings with red roofs, but this is speculative given the limited information available from such a small section of the image.

Figure 15: A sample case of Object classification.

Object color

Question: Determine the color of the object based on the given reference boundary box. The resolution of the satellite image is 7360 x 4912. Bounding box coordinates: [3620, 1651, 3604, 1759].

Options
(A) Red
(B) Yellow
(C) Black
(D) White

Answer: C

4912×7360

LLaVA-Next: C

The color of the object based on the given reference boundary box is (C) Black. The object in question appears to be a building or a section of a building with a dark roof. The image resolution is 7360 x 4912, which allows for a clear view of the object's color. The bounding box coordinates [3620, 1651, 3604, 1759] indicate the specific area of the image where the object is located. The object's color is consistent with the color of the roof, which is black. The surrounding area includes other buildings, roads, and green spaces, but the object in question is the only one that matches the color (C) Black.

Perception Error

Qwen2-VL:A

The correct answer is (A) Red. The object in question is located at the coordinates [3620, 1651, 3604, 1759] in the satellite image. Based on the color and shape of the object, it appears to be a tennis court, which is typically colored red. The surrounding area includes buildings and streets, which are not the focus of this question. Therefore, the correct answer is (A) Red.

Perception Error

LLaVA-OneVision: A

The color of the object in question is red. This can be determined by examining the bounding box coordinates provided, which are [3620, 1651, 3604, 1759]. By looking at the image and focusing on the area defined by these coordinates, it becomes clear that the object has a distinct red color. The surrounding objects have different colors, such as white, gray, and blue, but the object in question stands out due to its red hue.

Figure 16: A sample case of Object color.

Object motion state

10000×10000

Question: Determine the motion state of an object based on the given reference bounding box. The resolution of the satellite image is 10000 x 10000. Bounding box coordinates: [7846, 428, 7862, 440].

Options
(A) Yes
(B) No

Answer: A

Perception Error

LLaVA-Next:B

The reference bounding box coordinates you provided are very small and do not correspond to any significant object in the image. The image appears to be a satellite view of a landscape with fields, roads, and possibly a body of water. The bounding box you provided is too small to be discernible or to determine the motion state of an object. To determine the motion state of an object, one would typically look for changes in position, orientation, or size over time, which are not apparent from the given bounding box coordinates.

Lack of Knowledge

Qwen2-VL:B

The given reference bounding box coordinates [7846, 428, 7862, 440] are within the image, but the object within this bounding box appears to be a small, isolated patch of land or water. There is no clear indication of motion, such as a moving object or a change in the position of the object over time. Therefore, based on the given image and the reference bounding box, it is not possible to determine the motion state of the object.

Lack of Knowledge

LLaVA-OneVision: B

The bounding box coordinates provided are [7846, 428, 7862, 440]. These coordinates represent a very small rectangular area within the larger image. Given that the resolution of the satellite image is 10000 x 10000, this bounding box corresponds to a very tiny portion of the entire image. To determine the motion state of an object based on this bounding box, we would need more information about what kind of object we are looking at and how it might move. However, since the bounding box is so small and does not encompass any significant features or objects, it is not possible to accurately determine the motion state of anything within this bounding box. Therefore, the answer to the question "Determine the motion state of an object based on the given reference bounding box" with the bounding box coordinates [7846, 428, 7862, 440] is B) No, because the bounding box is too small to provide any meaningful information about motion.

Figure 17: A sample case of Object motion state.

Object spatial relationship

Question: The upper right corner of the picture is an open area.
In which direction is the tree facing the person

Options
(A) Upper Right
(B) Above
(C) Below
(D) Lower Left

Answer: D

4912×7360

Perception Error

LLaVA-Next:A

The tree in the image is facing the person in the direction of the upper right corner. This is determined by the orientation of the tree's foliage and the direction of the shadow it casts. The tree appears to be well-lit from the upper right, which suggests that the sun is shining on it from that direction. The shadow of the tree is cast to the left, indicating that the tree is facing towards the upper right corner of the image.

Perception Error

Qwen2-VL:A

The correct answer is (A) Upper Right. The tree in the upper right corner of the picture is facing the person. This can be determined by observing the direction of the tree's branches and leaves, which are oriented towards the upper right part of the image.

LLaVA-OneVision: D
The tree is facing the person in the lower left corner of the image.

Figure 18: A sample case of Object spatial relationship.

7 Datasheets

In this section, we document essential details about the proposed datasets and benchmarks following the CVPR Dataset and Benchmark guidelines and the template provided by Gebru *et al.* [8].

7.1 Motivation

The questions in this section are primarily intended to encourage dataset creators to clearly articulate their reasons for creating the dataset and to promote transparency about funding interests. The latter may be particularly relevant for datasets created for research purposes.

1. *"For what purpose was the dataset created?"*

   A: Existing benchmarks often use significantly smaller image sizes than those found in real-world RS scenarios, have limited annotation quality, and fail to account for key evaluation dimensions. To address these limitations, we introduce XLRS-Bench, a comprehensive benchmark designed to assess the perception and reasoning capabilities of MLLMs in ultra-high-resolution RS contexts.

2. *"Who created the dataset (e.g., which team, research group) and on behalf of which entity?"*

   A: The authors of this anonymous CVPR submission.

3. *"Who funded the creation of the dataset?"*

   A: The dataset creation was funded by the affiliations of the authors involved in this work.

7.2 Composition

Most of the questions in this section are intended to provide dataset consumers with the information they need to make informed decisions about using the dataset for their chosen tasks. Some of the questions are designed to elicit information about compliance with the EU's General Data Protection Regulation (GDPR) or comparable regulations in other jurisdictions. Questions that apply only to datasets that relate to people are grouped together at the end of the section. We recommend taking a broad interpretation of whether a dataset relates to people. For example, any dataset containing text that was written by people relates to people.

1. *"What do the instances that comprise our datasets represent (e.g., documents, photos, people, countries)?"*

   A: The dataset primarily consists of ultra-high-resolution remote sensing images captured by satellites, along with their corresponding textual annotations. All datasets utilized in XLRS-Bench are publicly accessible and nonprofit.

2. *"How many instances are there in total (of each type, if appropriate)?"*

   A: XLRS-Bench includes 1,400 ultra-high-resolution images, with 840 reaching a resolution of 10,000 $\times$ 10,000. Additionally, for these ultra-high-resolution images, we have provided 934 detailed captions, 32,389 VQA pairs, and 12,619 visual grounding instances.

3. *"Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set?"*

   A: The images in XLRS-Bench are sourced from existing detection [9, 10] and segmentation [11, 12] datasets, but all textual annotations were independently created by us.

4. *"Is there a label or target associated with each instance?"*

   A: Yes, for these ultra-high-resolution images, we have provided 934 detailed captions, 32,389 VQA pairs, and 12,619 visual grounding instances.

5. *"Is any information missing from individual instances?"*

   A: No, each individual instance is complete.

6. *"Are relationships between individual instances made explicit (e.g., users' movie ratings, social network links)?"*

   A: Yes, the relationship between individual instances is explicit.

7. *"Are there recommended data splits (e.g., training, development/validation, testing)?"*

   A: The dataset is designed to evaluate the perception and reasoning abilities of MLLMs, so we recommend using it in its entirety as a test set.

8. *"Is the dataset self-contained, or does it link to or otherwise rely on external resources (e.g., websites, tweets, other datasets)?"*

   A: XLRS-Bench is self-contained and will be open-sourced on platforms like Hugging Face, integrated into evaluation tools such as LLMs-Eval [13, 14] for easy use.

9. *"Does the dataset contain data that might be considered confidential (e.g., data that is protected by legal privilege or by doctor–patient confidentiality, data that includes the content of individuals' non-public communications)?"*

   A: No, all data are clearly licensed.

10. *"Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety?"*

    A: No, XLRS-Bench does not contain any data with negative information.

7.3 Collection Process

In addition to the goals outlined in the previous section, the questions in this section are designed to elicit information that may help researchers and practitioners create alternative datasets with similar characteristics. Again, questions that apply only to datasets that relate to people are grouped together at the end of the section.

1. *"How was the data associated with each instance acquired?"*

   A: The images in XLRS-Bench are sourced from existing detection [9, 10] and segmentation [11, 12] datasets. We enrich these ultra-high-resolution images with manual annotations, including 934 detailed captions, 32,389 VQA pairs, and 12,619 visual grounding instances.

2. *"What mechanisms or procedures were used to collect the data (e.g., hardware apparatuses or sensors, manual human curation, software programs, software APIs)?"*

   A: We employed professional annotation and quality control teams to complete the annotations for VQA and Visual Grounding tasks. For the Image Captioning task, we developed a semi-automated pipeline. Detailed information can be found in Section 3.2 of the main text.

3. *"If the dataset is a sample from a larger set, what was the sampling strategy (e.g., deterministic, probabilistic with specific sampling probabilities)?"*

   A: Please refer to the details listed in the main text Section 3.2.

7.4 Preprocessing, Cleaning, and Labeling

The questions in this section are intended to provide dataset consumers with the information they need to determine whether the "raw" data has been processed in ways that are compatible with their chosen tasks. For example, text that has been converted into a "bag-of-words" is not suitable for tasks involving word order.

1. *"Was any preprocessing/cleaning/labeling of the data done (e.g., discretization or bucketing, tokenization, part-of-speech tagging, SIFT feature extraction, removal of instances, processing of missing values)?"*

   A: Yes. During image collection, we prioritized selecting valuable satellite images for annotation. For linguistic annotation, three Level-3 subtasks—Regional Land Use Classification, Regional Counting, and Regional Counting with Change Detection—were marked with red circles. This method, mimicking human interaction, was essential for providing clear, fine-grained region-level analysis on ultra-high-resolution images.

2. *"Was the 'raw' data saved in addition to the preprocessed/cleaned/labeled data (e.g., to support unanticipated future uses)?"*

   A: Yes, raw data is accessible.

3. *"Is the software that was used to preprocess/clean/label the data available?"*

   A: Yes, the necessary software used to preprocess and clean the data is publicly available.

7.5 Uses

The questions in this section are intended to encourage dataset creators to reflect on tasks for which the dataset should and should not be used. By explicitly highlighting these tasks, dataset creators can help dataset consumers make informed decisions, thereby avoiding potential risks or harms.

1. *"Has the dataset been used for any tasks already?"*

   A: No.

2. *"Is there a repository that links to any or all papers or systems that use the dataset?"*

   A: Yes, we will provide such links in the GitHub and the Huggingface repository.

3. *"What (other) tasks could the dataset be used for?"*

   A: XLRS-Bench provides extensive annotations for VQA, Grounding, and Captioning tasks. In addition to evaluating the perception and reasoning capabilities of existing MLLMs, it can also be used to assess models specifically designed for these tasks.

4. *"Is there anything about the composition of the dataset or the way it was collected and preprocessed/cleaned/labeled that might impact future uses?"*

   A: No.

5. *"Are there tasks for which the dataset should not be used?"*

   A: N/A.

7.6 Distribution

Dataset creators should provide answers to these questions prior to distributing the dataset either internally within the entity on behalf of which the dataset was created or externally to third parties.

1. *"Will the dataset be distributed to third parties outside of the entity (e.g., company, institution, organization) on behalf of which the dataset was created?"*

   A: No. The datasets will be made publicly accessible to the research community.

2. *"How will the dataset be distributed (e.g., tarball on website, API, GitHub)?"*

   A: We will provide XLRS-Bench in the GitHub and the Huggingface repository.

3. *"When will the dataset be distributed?"*

   A: We will create a repository to release the data once the paper is officially published, ensuring compliance with the anonymity principle.

4. *"Will the dataset be distributed under a copyright or other intellectual property (IP) license, and/or under applicable terms of use (ToU)?"*

   A: Yes, the dataset will be released under the Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License.

5. *"Have any third parties imposed IP-based or other restrictions on the data associated with the instances?"*

   A: No.

6. *"Do any export controls or other regulatory restrictions apply to the dataset or to individual instances?"*

   A: No.

7.7 Maintenance

As with the questions in the previous section, dataset creators should provide answers to these questions prior to distributing the dataset. The questions in this section are intended to encourage dataset creators to plan for dataset maintenance and communicate this plan to dataset consumers.

1. *"Who will be supporting/hosting/maintaining the dataset?"*

   A: The authors of this work serve to support, host, and maintain the datasets.
2. *"How can the owner/curator/manager of the dataset be contacted (e.g., email address)?"*

   A: The curators can be contacted via the email addresses listed on our paper or webpage.
3. *"Is there an erratum?"*

   A: There is no explicit erratum; updates and known errors will be specified in future versions.
4. *"Will the dataset be updated (e.g., to correct labeling errors, add new instances, delete instances)?"*

   A: Future updates (if any) will be posted on the dataset website.
5. *"Will older versions of the dataset continue to be supported/hosted/maintained?"*

   A: Yes. This initial release will be updated in the future, with older versions replaced as new updates are posted.
6. *"If others want to extend/augment/build on/contribute to the dataset, is there a mechanism for them to do so?"*

   A: Yes, we will provide detailed instructions for future extensions.

8 Limitation and Potential Societal Impact

In this section, we discuss the limitations and potential societal impact of this work.

8.1 Potential Limitations

While **XLRS-Bench** provides a comprehensive benchmark for evaluating the perception and reasoning capabilities of MLLMs, there are several limitations to consider:

- **Scope of Sensors:** Although our benchmark includes 1,400 utlra-high-resolution visible light remote sensing images, it may not cover all possible real-world scenarios. There could be additional sensor data, like multispectral data that were not included in this study, potentially limiting the generalizability of our findings.
- **Model and Dataset Diversity:** In this paper, we extensively evaluated both general-purpose and RS-specific MLLMs. As new models emerge, their evaluation results will be added to our open-source leaderboard. Additionally, XLRS-Bench will also be expanded in dataset size and task diversity.
- **Multilingual Support:** XLRS-Bench currently supports both Chinese and English, surpassing the single-language limitations of most remote sensing benchmarks [15]. In the future, we aim to extend support to languages like Spanish and French.

8.2 Potential Negative Societal Impact

- **Safety Risks:** XLRS-Bench is designed to evaluate the performance of vision-language multimodal models in ultra-high-resolution remote sensing scenarios. However, excessive reliance on evaluation datasets may lead to overconfidence in autonomous systems, such as multimodal large models. It is crucial to implement adequate safety measures and human supervision when deploying these MLLMs to ensure public safety.

- **Environmental Impact:** Training MLLMs on large datasets and evaluating them using XLRS-Bench requires a certain amount of computational resources. To facilitate future research, we will maintain a leaderboard of MLLMs, removing the need for repeated evaluations of existing models.
- **Bias and Fairness:** XLRS-Bench, with its 16 Level-3 capabilities, is tailored for evaluating ultra-high-resolution remote sensing scenarios. However, it remains limited in comprehensiveness and may exhibit biases. For instance, disaster prediction in anomaly reasoning relies solely on satellite imagery, providing warnings but reflecting inherent biases. Effective decision-making demands the integration of local meteorological and hydrological data. In the future, we aim to expand the evaluation dimensions and datasets to deepen insights into ultra-high-resolution remote sensing applications.

References

[1] Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, et al. Mmbench: Is your multi-modal model an all-around player? In *European conference on computer vision*, pages 216–233. Springer, 2024.

[2] Bohao Li, Rui Wang, Guangzhi Wang, Yuying Ge, Yixiao Ge, and Ying Shan. Seed-bench: Benchmarking multimodal llms with generative comprehension. *arXiv preprint arXiv:2307.16125*, 2023.

[3] Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. Towards vqa models that can read. In *CVPR*, 2019.

[4] Ahmed Masry, Do Xuan Long, Jia Qing Tan, Shafiq Joty, and Enamul Hoque. Chartqa: A benchmark for question answering about charts with visual and logical reasoning. *arXiv preprint arXiv:2203.10244*, 2022.

[5] 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*, 2024.

[6] Kartik Kuckreja, Muhammad Sohail Danish, Muzammal Naseer, Abhijit Das, Salman Khan, and Fahad Shahbaz Khan. Geochat: Grounded large vision-language model for remote sensing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 27831–27840, 2024.

[7] Kaining Ying, Fanqing Meng, Jin Wang, Zhiqian Li, Han Lin, Yue Yang, Hao Zhang, Wenbo Zhang, Yuqi Lin, Shuo Liu, et al. Mmt-bench: A comprehensive multimodal benchmark for evaluating large vision-language models towards multitask agi. In *International Conference on Machine Learning*, pages 57116–57198. PMLR, 2024.

[8] Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé Iii, and Kate Crawford. Datasheets for datasets. *Communications of the ACM*, 64(12):86–92, 2021.

[9] Gui-Song Xia, Xiang Bai, Jian Ding, Zhen Zhu, Serge Belongie, Jiebo Luo, Mihai Datcu, Marcello Pelillo, and Liangpei Zhang. Dota: A large-scale dataset for object detection in aerial images. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3974–3983, 2018.

[10] Michael Ying Yang, Wentong Liao, Xinbo Li, and Bodo Rosenhahn. Deep learning for vehicle detection in aerial images. In *2018 25th IEEE International Conference on Image Processing (ICIP)*, pages 3079–3083. IEEE, 2018.

[11] Javiera Castillo-Navarro, Bertrand Le Saux, Alexandre Boulch, Nicolas Audebert, and Sébastien Lefèvre. Semi-supervised semantic segmentation in earth observation: The minifrance suite, dataset analysis and multi-task network study. *Machine Learning*, 111(9):3125–3160, 2022.

[12] Rodrigo Caye Daudt, Bertrand Le Saux, Alexandre Boulch, and Yann Gousseau. Multitask learning for large-scale semantic change detection. *Computer Vision and Image Understanding*, 187:102783, 2019.

[13] Kaichen Zhang, Bo Li, Peiyuan Zhang, Fanyi Pu, Joshua Adrian Cahyono, Kairui Hu, Shuai Liu, Yuanhan Zhang, Jingkang Yang, Chunyuan Li, and Ziwei Liu. Lmms-eval: Reality check on the evaluation of large multimodal models. *arXiv preprint arXiv:2407.12772*, 2024.

[14] Bo Li, Peiyuan Zhang, Kaichen Zhang, Fanyi Pu, Xinrun Du, Yuhao Dong, Haotian Liu, Yuanhan Zhang, Ge Zhang, Chunyuan Li, and Ziwei Liu. Lmms-eval: Accelerating the development of large multimodal models. `https://github.com/EvolvingLMMs-Lab/lmms-eval`, March 2024.

[15] Xiang Li, Jian Ding, and Mohamed Elhoseiny. Vrsbench: A versatile vision-language benchmark dataset for remote sensing image understanding. *arXiv preprint arXiv:2406.12384*, 2024.