VRSBench: A Versatile Vision-Language Benchmark Dataset for Remote Sensing Image Understanding

Xiang Li
KAUST

Jing Ding
KAUST

Mohamed Elhoseiny
KAUST

Background

- Existing vision-language datasets primarily cater to single image perception tasks.
- Mostly provide only brief descriptions, lacking detailed object information.
- Current remote sensing visual grounding datasets are designed under simplistic scenarios where the referring objects typically stand alone within their category.
- Current remote sensing VQA dataset contains close-set answers.

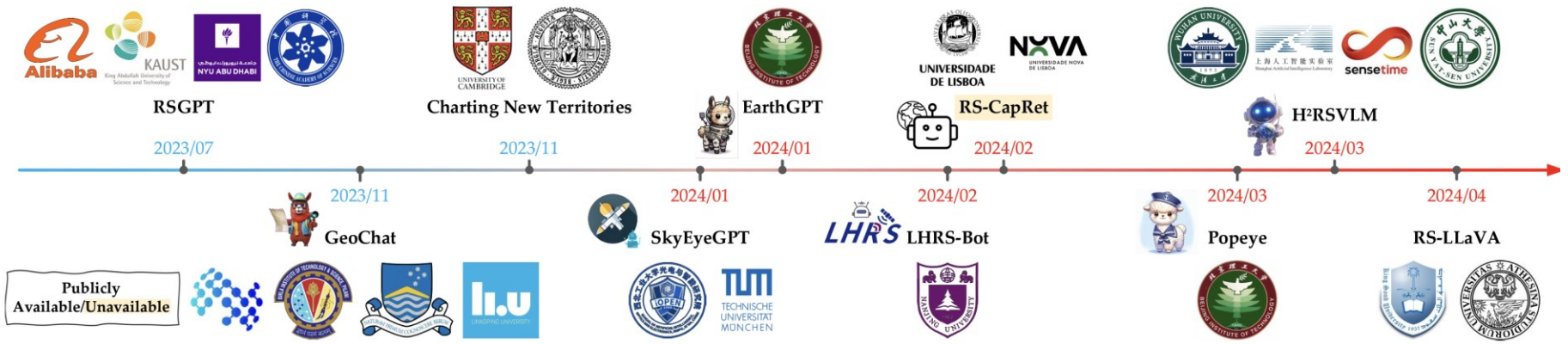


Fig 1. The timeline of recent Large Vision-Language Modes for Remote Sensing

https://github.com/ZhanYang-nwpu/Awesome-Remote-Sensing-Multimodal-Large-Language-Model

Motivation

➢Our contributions

Dataset	Year	#Image	Caption		Grounding		VQA		Human
			#Captions	Details	#Refers	OBB	#VQAs	Open-set	
UCM-Captions [30]	2016	2,100	10,500 (12)	✗	0	✗	0	✗	✓
RSICD [10]	2017	10,921	54605 (12)	✗	0	✗	0	✗	✓
RS5M [31]	2023	5M	5M (49)	✓	0	✗	0	✗	✗
RSICap [27]	2023	2,585	2,585 (60)	✓	0	✗	0	✗	✓
RSVG [18]	2022	4,239	0	✗	7,933	✗	0	✗	✓
DIOR-RSVG [19]	2023	17,402	0	✗	38,320	✗	0	✗	✓
RRSIS-D [20]	2024	17,402	0	✗	17,402	✗	0	✗	✓
RSVQA-HR [21]	2020	10,659	0	✗	0	✗	1,066,316	✗	✗
RSIVQA [22]	2021	37,264	0	✗	0	✗	111,134	✓	✓
VQA-TextRS [24]	2022	2,144	0	✗	0	✗	6,245	✓	✓
RSIEval [27]	2023	100	0	✗	0	✗	933	✓	✓
VRSBench	-	29,614	29,614 (52)	✓	52,472	✓	123,221	✓	✓

Table 1. Comparison between existing remote sensing vision-language datasets and VRSBench dataset.

Key features: **1) a large-scale collection of human-verified, high-quality captions rich in object details; 2) more realistic object refers; 3) diverse open-set question-answer pairs in natural language.**

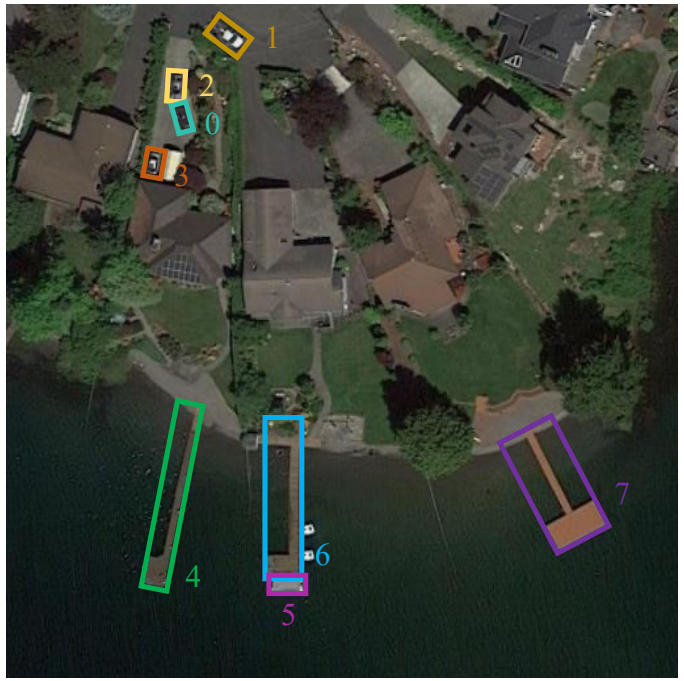
VRSBench

➢Dataset Summary

- A versatile dataset with human-verified detailed caption, complex object referring (visual grounding) and visual question answering (VQA).
  - Caption: 29,614 detailed captions.
  - Grounding: 52,472 complex object referring.
  - Complex VQA: 123,221 open-set question-answer pairs.

VRSBench

➢Example



Object Referring

Object ID=1: The small vehicle that is the farthest to the top.
Object ID=4: The harbor located on the left side of the scene with multiple docks extending into the water.
Object ID=7: The harbor situated on the right side of the image with a large dock area.

Visual Question Answer

Question: How many harbors are visible? *Answer:* 3
Question: What is the object located furthest to the top? *Answer:* Small vehicle
Question: Are the visible vehicles near water? *Answer:* No

Detailed Captioning

The high-resolution aerial image from GoogleEarth shows a waterfront scene with residential areas and harbor facilities. Three distinct harbors can be seen, one located on the left side and another on the right side of the image. Between them, there are homes with different colored rooftops, green lawns, and driveways. A ship is docked in the central part of the bottom edge, and the water body exhibits gentle ripples. Various small vehicles are scattered throughout the residential area, parked near the houses.

VRSBench Data Collection

➢Data Collection Pipeline

Attribute Extraction

Extracted image and object information: {'source': image source, 'resolution': image resolution, 'objects': [{'obj_id': object id, 'obj_cls': object category, 'obj_bbox': bounding box, 'is_unique': unique within category, 'obj_position': object position, 'obj_size': object size, ...}]}

Prompt Engineering

Instructions: you are an AI visual assistant tasked with analyzing remote sensing images. Given an input image and object information. Your job is to create a detailed image caption, referring sentences for distinct objects, as well as question-answer pairs…

GPT-4 Inference

Inputs: image, image and object information in JSON format.
Outputs: {'caption', detailed image caption, 'objects': [{obj_id, referring_sentence},...], 'qa_pairs': ['ques_id': question id, 'question': question, 'type': question type, 'answer': answer]}

Human Verification

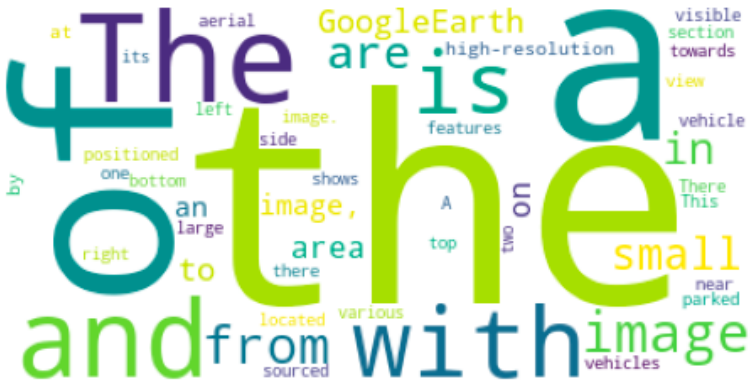
Guidelines: 1) modify uncertain or meaningless elements; 2) modify error referring; 2) remove self-answering questions…
Outputs: corrected annotations.

VRSBench Statistics

➢Image Caption

#images	29,614
#vocabulary size	9,588
#total words	1,526,338
#caption sentences	114,366
Avg. #sentences in caption	4
Avg. caption length	52

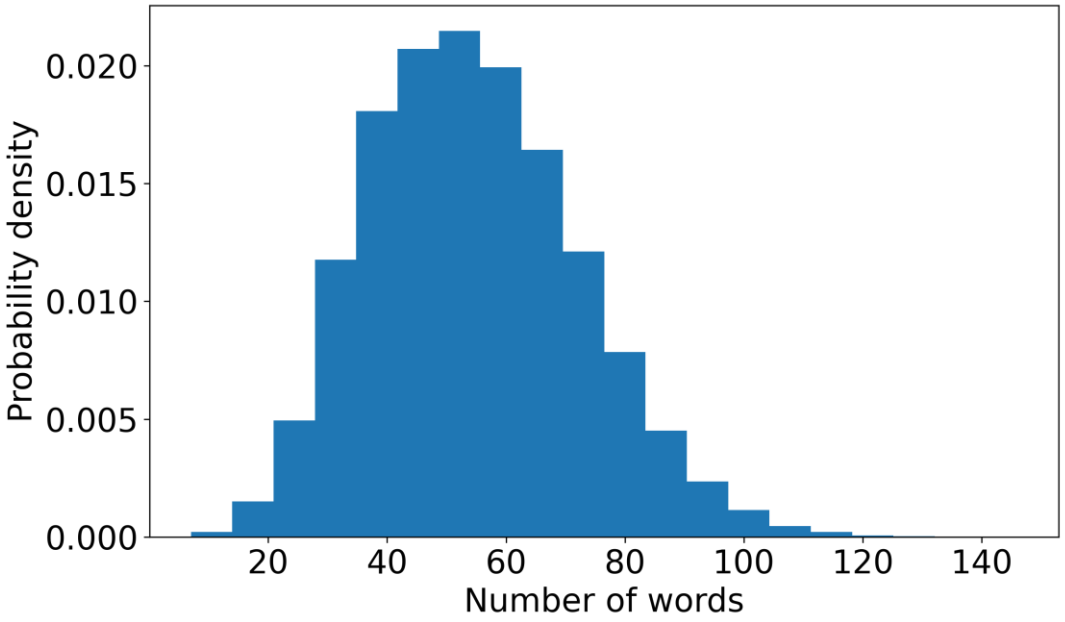
(a) Caption Summary.



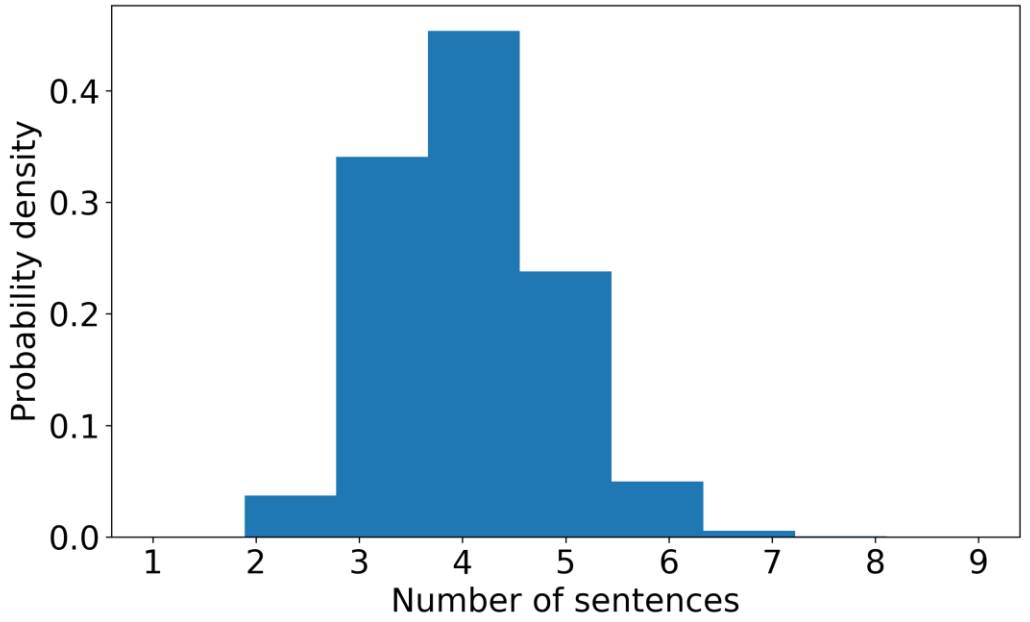
(b) Word cloud of captions.

VRSBench Statistics

- Image Caption



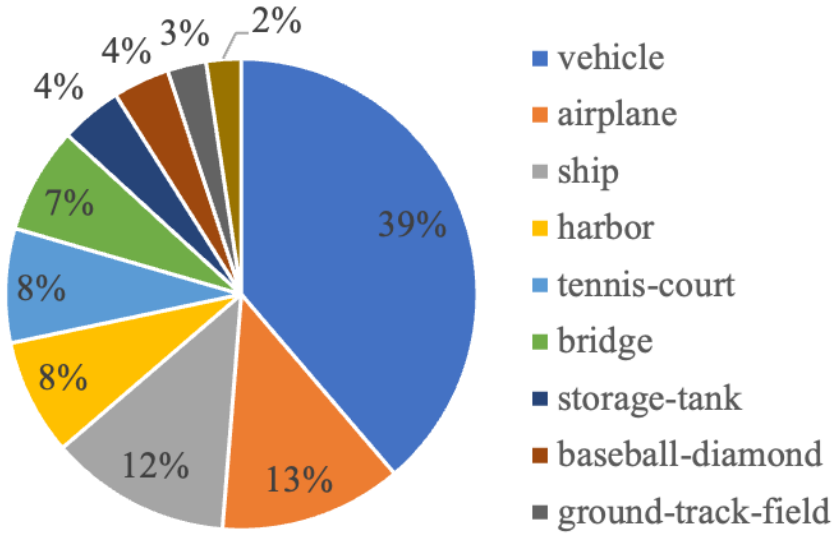
(c) Distribution of caption word length.



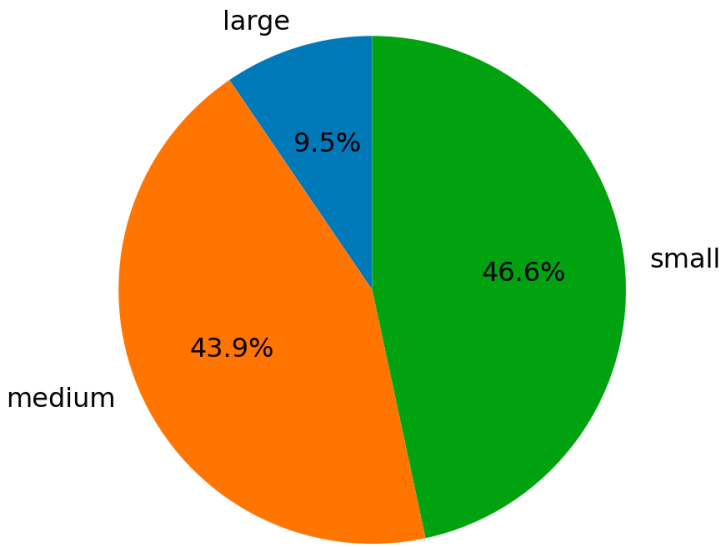
(d) Distribution of caption sentence length.

VRSBench Statistics

➢Visual Grounding



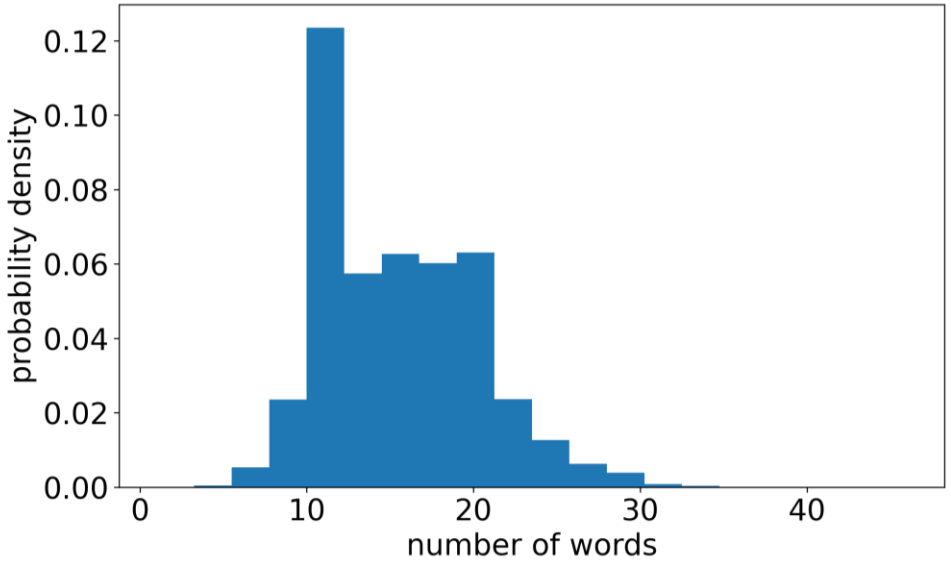
(a) Distribution of top-10 object category.



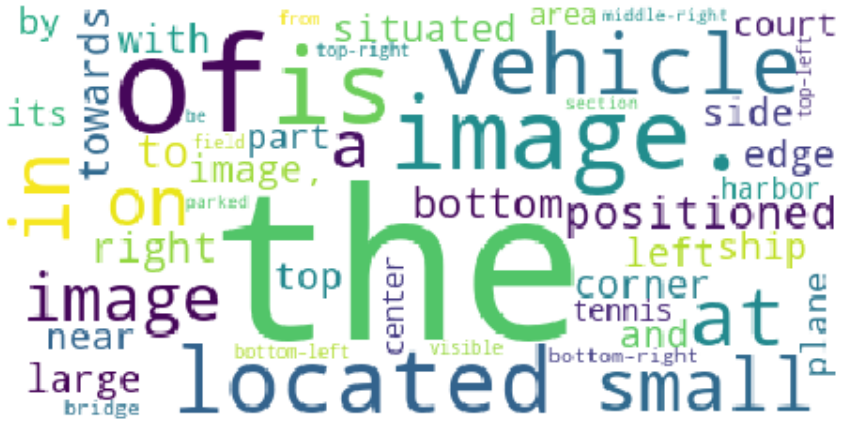
(b) Distribution of object size.

VRSBench Statistics

➢Visual Grounding



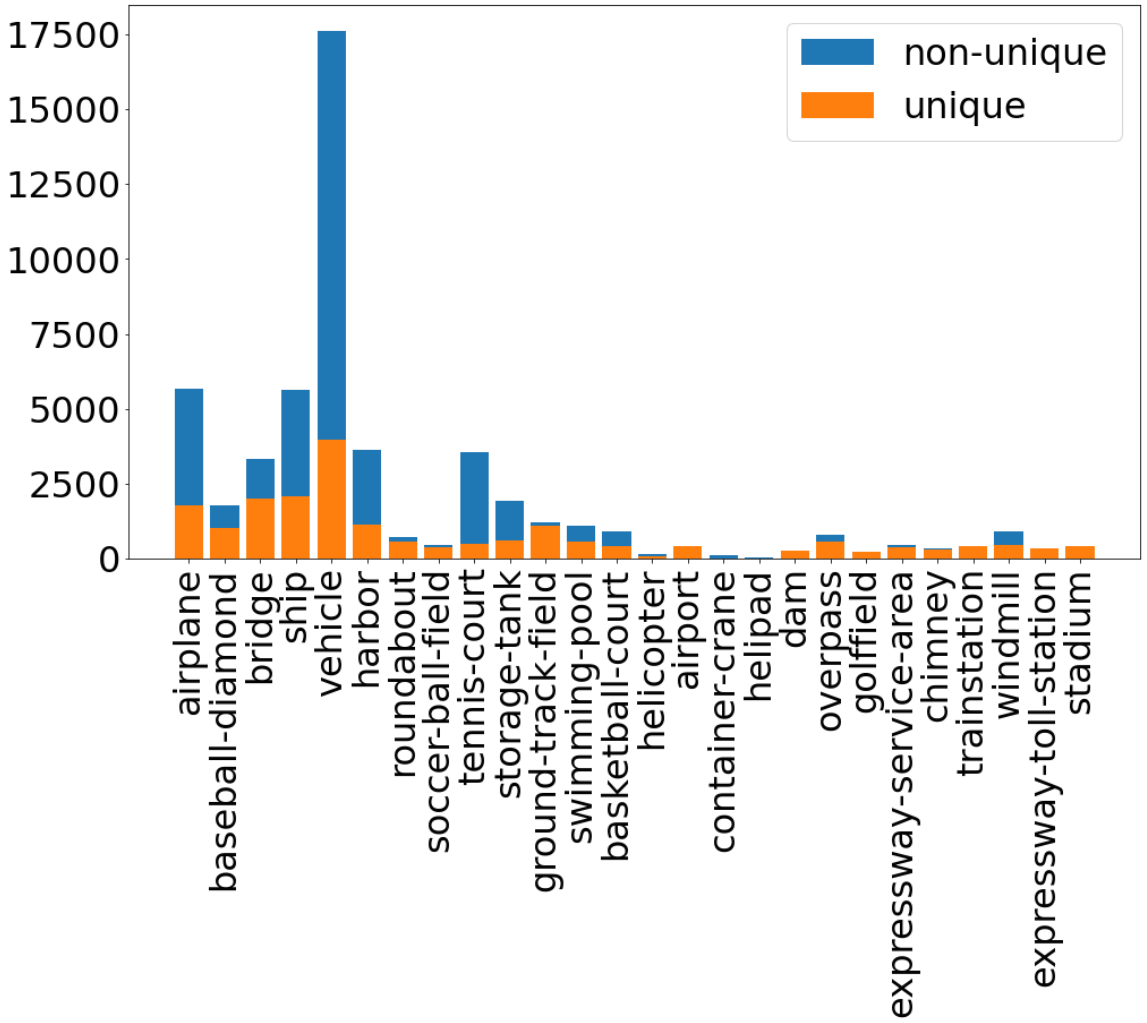
(c) Distribution of word length.



(d) Word cloud of referring sentences.

VRSBench Statistics

➢Visual Grounding



(e) Distribution of unique/non-unique referring objects per category.

VRSBench Statistics

➢Visual Question Answering

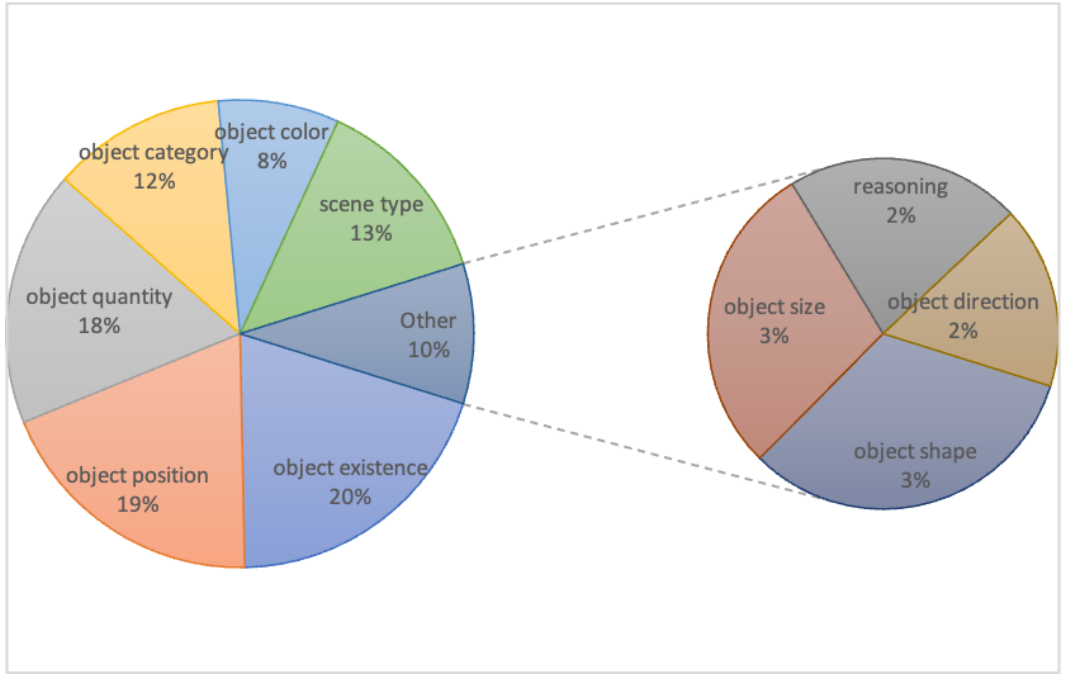


Fig 1. Distribution of question type.

Fig 2. Word cloud of questions.

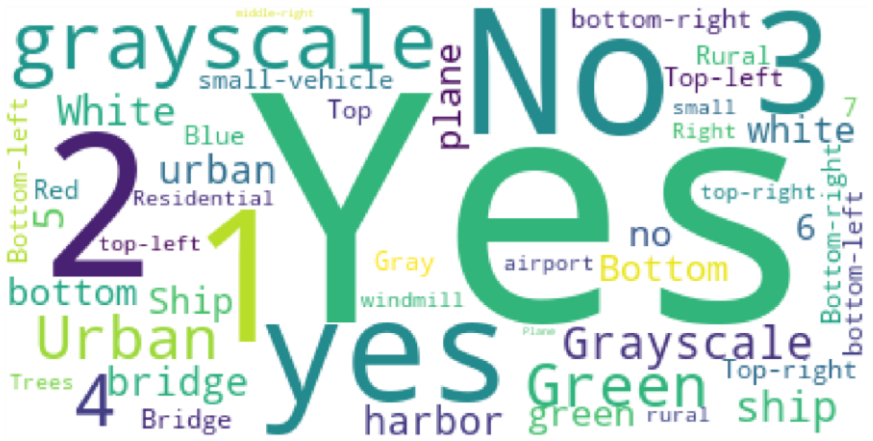


Fig 3. Word cloud of answers.

VRSBench Benchmark

➢Benchmark Tasks

- VRSBench-Cap: This challenge requires the prediction of a comprehensive description for a given remote sensing image, encapsulating intricate details and contextual relevance.
- VRSBench-Ref: The task involves identifying and localizing specific objects or features within a given remote sensing image based on a textual description.
- VRSBench-VQA: Participants are prompted to answer questions related to visual content in a given remote sensing image.

VRSBench Benchmark

➢Benchmark Settings

- We benchmark state-of-the-art models, including LLaVA-1.5, MiniGPT-v2, and GeoChat, to demonstrate the potential of LVMs for remote sensing image understanding.
- We reload the models that are initially trained on large-scale image-text alignment datasets, and then finetune each method using the training set of our VRSBench dataset for 5 epochs.
- We employ CLIP-ViT(L-14) as the vision encoder and use the Vicuna-7B model as the LLM.

VRSBench Benchmark

➢Data Splits

We split the datasets according to official splits of DOTA and DIOR datasets, where their training images are used to build the training set of VRSBench and their validation sets are used as the test set.

	train	test
#Images	20,264	9,350
#Captions	20,264	9,350
#Refers	36,313	16,159
#VQAs	85,813	37,408

Table 2: VRSBench Data Splits

VRSBench Benchmark

➢Detailed Image Captioning

- BLEU, ROUGE_L, METEOR, and CIDEr for caption evaluation.
- We also report GPT-based CHIAR score.

Method	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE_L	CIDEr	CHAIR	Avg_L
GeoChat w/o ft [28]	13.9	6.6	3.0	1.4	7.8	13.2	0.4	0.42	36
GPT-4V [32]	37.2	22.5	13.7	8.6	20.9	30.1	19.1	**0.83**	67
MiniGPT-v2 [38]	36.8	22.4	13.9	8.7	17.1	30.8	21.4	0.73	37
LLaVA-1.5 [37]	**48.1**	**31.5**	**21.2**	**14.7**	**21.9**	**36.9**	**33.9**	0.78	49
GeoChat [28]	46.7	30.2	20.1	13.8	21.1	35.2	28.2	0.77	52
Mini-Gemini [48]	47.6	31.1	20.9	14.3	21.5	36.8	33.5	0.77	47

Table 3: Detailed image caption performance on VRSBench dataset. Avg_L denotes the average word length of generated captions.

VRSBench Benchmark

➢Visual Grounding

- For model evaluation, we employ the metric accuracy@**τ** to assess performance.

Method	Unique		Non Unique		All	
	Acc@0.5	Acc@0.7	Acc@0.5	Acc@0.7	Acc@0.5	Acc@0.7
GeoChat w/o ft [28]	20.7	5.4	7.3	1.7	12.9	3.2
GPT-4V [32]	8.6	2.2	2.5	0.4	5.1	1.1
MiniGPT-v2 [38]	40.7	18.9	32.4	15.2	35.8	16.8
LLaVA-1.5 [37]	51.1	16.4	34.8	11.5	41.6	13.6
GeoChat [28]	**57.4**	**22.6**	**44.5**	**18.0**	**49.8**	**19.9**
Mini-Gemini [48]	41.1	9.6	22.3	4.9	30.1	6.8

Table 4. Visual grounding performance on VRSBench dataset.

VRSBench Benchmark

➢Visual Question Answering

- We categorized the questions in the test set into 10 distinct types: object category, presence, quantity, color, shape, size, position, direction, scene characteristic, and reasoning.
- We use a GPT-based protocol for open-set VQA evaluation.

Method	Category	Presence	Quantity	Color	Shape	Size	Position	Direction	Scene	Reasoning	All
# VQAs	5435	7789	6374	3550	1422	1011	5829	477	4620	902	
GeoChat w/o ft [28]	48.5	85.9	19.2	17.0	18.3	32.0	43.4	42.1	44.2	57.4	40.8
GPT-4V [32]	67.0	87.6	45.6	71.0	70.8	54.3	67.2	50.7	69.8	72.4	65.6
MiniGPT-v2 [38]	61.3	26.0	46.1	51.0	41.8	11.2	17.1	12.4	49.3	21.9	38.2
LLaVA-1.5 [37]	86.9	91.8	58.2	69.9	72.2	**61.5**	**69.5**	**56.7**	**83.9**	73.4	76.4
GeoChat [28]	86.5	**92.1**	56.3	70.1	73.8	60.4	69.3	53.5	83.7	73.5	76.0
Mini-Gemini [48]	**87.8**	92.1	**58.8**	**74.0**	**75.3**	58.0	68.0	56.7	83.2	**74.4**	**77.8**

Table 5. Visual question answering performance on VRSBench dataset

Thanks for your attention!