RemoteSAM: Towards Segment Anything for Earth Observation

Liang Yao[1,*], **Fan Liu**[1,*,†], **Delong Chen**[2,*], **Chuanyi Zhang**[1]
Yijun Wang[1], **Ziyun Chen**[1], **Wei Xu**[1], **Shimin Di**[3], **Yuhui Zheng**[1]

[1]Hohai University [2]HKUST [3]Southeast University

[*]Equal Contribution [†]Corresponding Author
Email: fanliu@hhu.edu.cn

Abstract

We aim to develop a robust yet flexible visual foundation model for Earth observation. It should possess strong capabilities in recognizing and localizing diverse visual targets while providing compatibility with various input-output interfaces required across different task scenarios. Current systems cannot meet these requirements, as they typically utilize task-specific architecture trained on narrow data domains with limited semantic coverage. Our study addresses these limitations from two aspects: data and modeling. We first introduce an automatic data engine that enjoys significantly better scalability compared to previous human annotation or rule-based approaches. It has enabled us to create the largest dataset of its kind to date, comprising 270K image-text-mask triplets covering an unprecedented range of diverse semantic categories and attribute specifications. Based on this data foundation, we further propose a task unification paradigm that centers around referring expression segmentation. It effectively handles a wide range of vision-centric perception tasks, including classification, detection, segmentation, grounding, etc, using a single model without any task-specific heads. Combining these innovations on data and modeling, we present RemoteSAM, a foundation model that establishes new SoTA on several earth observation perception benchmarks, outperforming other foundation models such as Falcon, GeoChat, and LHRS-Bot with significantly higher efficiency. Models and data are publicly available at https://github.com/1e12Leon/RemoteSAM.

1. Introduction

Advances in AI have fundamentally transformed Earth observation paradigms [62]. Strong visual perception models are driving breakthroughs across diverse applications, such as urban development [8], agriculture [63], disaster management [1], and beyond. As these tasks often involve distinct input-output interfaces, they are typically handled individually by specialized models. To effectively manage this heterogeneity, we are interested in developing a foundational model [2, 82] that unifies multiple perception tasks. Such a model would conveniently accommodate varied application scenarios and effectively integrate knowledge learned across different tasks and domains.

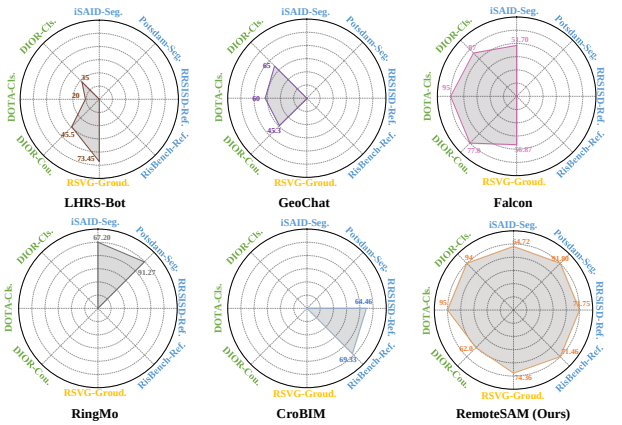


Figure 1. Comparison of various foundation models for Earth observation [17, 25, 50, 59, 76]. Blue, yellow, and green represent pixel-level, region-level, and image-level tasks, respectively. Our RemoteSAM is competitive with other models on most datasets and performs significantly better on pixel-level tasks.

Current attempts at task unification mainly fall into two paradigms, as represented in Fig. 2. Methods based on **task-specific heads** (*e.g.*, Embedding Fields [4], RingMo [59], ScaleMAE [53], and various extensions of RemoteCLIP models [32, 38, 85]) integrate general-purpose encoders with specialized decoders tailored to individual downstream tasks. However, the knowledge sharing between differ-

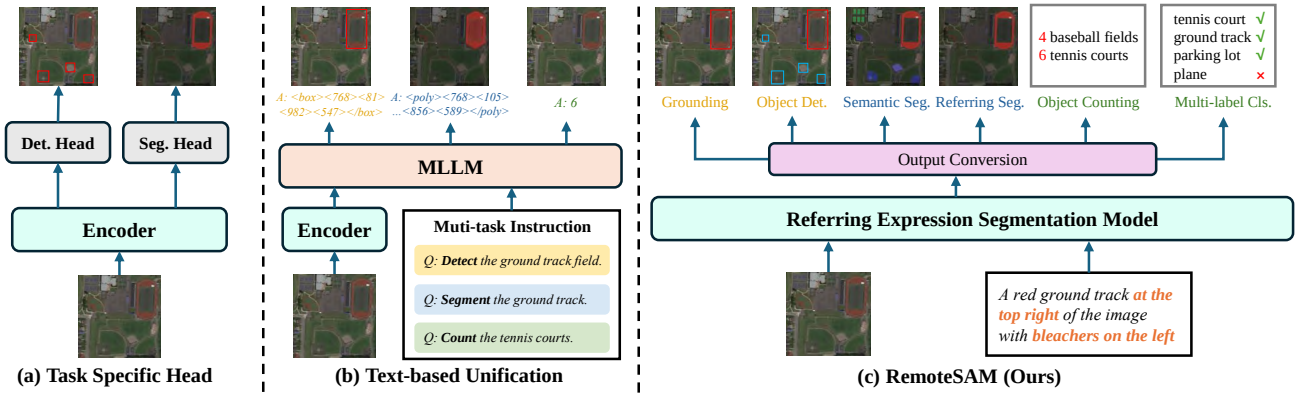


Figure 2. Comparison of different foundation models of remote sensing. (a) Task-specific head methods [14, 33]. (b) Text-based unification approaches [25, 76]. (c) Our proposed referring segmentation-based paradigm, RemoteSAM. It is a more unified architecture with fewer parameters, achieving enhanced pixel-level understanding performance.

ent task heads is inherently limited, and it necessitates task-specific fine-tuning whenever new objectives emerge. Other methods, exemplified by Falcon [76], GeoChat [25], employs **text-based task unification** with language models [9, 10, 68]. Although achieving good performance in image-level (*e.g.*, scene classification) and region-level tasks (*e.g.*, object detection), such approaches exhibit intrinsic limitations in pixel-level prediction tasks, as natural language is not suitable for representing dense pixel-level outputs.

These limitations motivate us to design a flexible foundation model architecture capable of seamlessly supporting heterogeneous input-output interfaces across multiple tasks and granularities. Our proposed model operates at the fundamental pixel-level granularity, which acts as the most fundamental and indivisible output unit, enabling seamless upward compatibility to region-level and image-level tasks. Additionally, to facilitate convenient adaptation, the model integrates natural language understanding capabilities. This approach resembles the Referring Expression Segmentation (RES) [11, 19, 24, 71] task, where a model generates a pixel-level map from the input image conditioned on a text prompt.

In this paper, we present **RemoteSAM**, a vision foundation model built upon a novel architectural paradigm centered on RES. Our model leverages the dense pixel-level outputs inherent in RES, effectively converting these outputs into various formats required by other vision-centric tasks. Unlike existing VLM-based foundation models such as Falcon [76], LHRS-Bot [50], EarthGPT [83], and GeoChat [25], our RemoteSAM seamlessly supports pixel-level (*e.g.*, segmentation), region-level (*e.g.*, grounding), and image-level (*e.g.*, counting) tasks within a unified architecture. Furthermore, by eliminating the large LLM backbone—which contributes minimally to visual perception—RemoteSAM achieves significant parameter efficiency, enabling it to efficiently process high-resolution remote sensing data.

To equip RemoteSAM with robust semantic understanding capabilities, an RES dataset with extensive semantic coverage is necessary. However, existing datasets [43, 79] suffer from limited categorical diversity, restricted attribute variation, and overly templated expressions. To overcome these limitations, we construct a semantically diverse dataset using a scalable automated data curation pipeline. Observing that recent VLMs [73] demonstrate strong image comprehension abilities, we leverage them to extract rich semantic information from remote sensing imagery, generating referring expressions with broad linguistic diversity. Through iterative pseudo-label refinement utilizing mixed teacher models, we create a comprehensive dataset containing 270K Image-Text-Mask triplets, named **RemoteSAM-270K**. It is characterized by two critical dimensions: 1) extensive category Scope, encompassing diverse and prevalent remote sensing targets; and 2) multifaceted attribute completeness, employing linguistically varied expressions to describe detailed object attributes such as colors, states, spatial relations, and other distinctive visual-semantic characteristics.

To quantitatively validate the semantic coverage of remote sensing datasets, we further establish a hierarchical remote sensing semantic vocabulary, named *RSVocab-1K*, comprising 1K prevalent object categories. The qualitative results based on RSVocab-1K validate that RemoteSAM-270K's referring expressions possess a diverse range of category completeness and attribute expressiveness. Benefiting from our high-quality dataset, the model achieves superior adaptability to more unseen categories and datasets. We also conduct holistic evaluations to measure the performance of RemoteSAM across multiple downstream visual-centric tasks of remote sensing. The experiment result demonstrates that RemoteSAM effectively addresses

the persistent performance gap of conventional foundation models in fine-grained pixel-wise prediction tasks with an order-of-magnitude smaller parameter count (from billions to millions). For example, RemoteSAM achieves significant performance improvements on referring expression segmentation, outperforming existing methods by more than 3.0% on both RRSISD and RisBench benchmarks in terms of mIoU. It also achieves state-of-the-art semantic segmentation performance without fine-tuning, surpassing vision foundation models, *e.g.*, MA3E [33] and Scale-MAE [54]. In addition, it yields a remarkable 35% accuracy gain over GeoChat [25] in multi-label classification with a parameter-efficient architecture.

The contributions of this paper to remote sensing are summarized as follows:

- We propose a robust yet flexible visual foundation model for Earth observation, RemoteSAM. To the best of our knowledge, it is the first exploration and practice of referring segmentation-based task unification paradigm.
- We build a new referring expression dataset with an automatic data curation pipeline. It significantly exceeds the scale of existing datasets and possesses rich semantic coverage, benefiting from the advantages of VLMs.
- We construct a hierarchical semantic vocabulary. It can be utilized to evaluate the semantic coverage of remote sensing datasets, which helps measure their ability to adapt to real-world applications.
- Holistic evaluations demonstrate that RemoteSAM's superior performance over existing approaches with prominent fewer parameters. It demonstrates remarkable performance across multiple visual-centric tasks, particularly in pixel-level interpretation tasks.

2. Related Work

2.1. Remote Sensing Foundation Models

Recent advances in foundation models [2, 40, 73] have catalyzed transformative progress across multiple computer vision domains. However, despite their success in processing natural images, existing VFMs exhibit critical limitations when applied to remote sensing imagery [39, 49, 77] — a domain characterized by multimodal signals (*e.g.*, multi-spectral bands), fine-grained spatial details (*e.g.*, sub-meter resolutions), complex geospatial relationships, and temporal dynamics across acquisition epochs. Therefore, researchers have initiated systematic development of remote sensing foundation models. Initially, RingMo [60] is the first to construct a generative self-supervised framework. Through multi-modal data augmentation and scene-aware contrastive learning, it solves the problem of representation in complex remote sensing scenes. To capture temporal evolution patterns, SatMAE [14] designs a temporal embedding encoding and cross-time mask reconstruction strategy, establishing a new paradigm for dynamic remote sensing sequence modeling. To address the bottleneck in multi-scale representation, Scale-MAE [53] proposes a scale-aware masking strategy and a hierarchical decoder architecture, achieving scale-invariant learning of geospatial features. BFM [5] constructs a large remote sensing model with billions of parameters for the first time through a mixture-of-experts architecture and distributed training technology, verifying the feasibility of model scaling. The recently proposed SatMAE++ [51] further integrates multi-resolution pre-training and a convolutional upsampling module, achieving hierarchical fusion of cross-scale features, thereby integrating multi-scale information and enhancing the modeling capability of remote sensing images. Inspired by MiniGPT-4 [88], GeoChat [25] achieves the ability of visual grounding by utilizing a novel multimodal instruction dataset. LHRS-Bot [50] creates LHRS-Instruct, using globally available remote sensing images and corresponding OpenStreetMap features to exhibit a profound comprehension of RS images.

2.2. Remote Sensing Referring Segmentation

Referring Image Segmentation [11, 24] aims to segment specific objects from natural images utilizing natural language descriptions. Existing natural image segmentation methods are difficult to apply directly to remote sensing images. LGCE [79] defines the Referring Remote Sensing Image Segmentation (RRSIS) task and provides the RefSegRS dataset. It also presents a language-guided cross-scale enhancement module designed to improve small object segmentation by integrating multi-scale visual and linguistic features. However, the limited data scale constrains LGCE's performance with complex data. Consequently, RMSIN [43] proposed the RRSIS-D dataset and addresses the diverse scales and orientations of objects in remote sensing images through intra-scale and cross-scale interaction modules as well as adaptive rotated convolution. To overcome the limitations of current RRSIS datasets—small size, single spatial resolution, and sparse object samples—CroBIM [17] introduces RISBench, offering more comprehensive and challenging data. They also design a cross-modal bidirectional interaction model with context-aware prompt modulation, which enhances segmentation performance in complex remote sensing backgrounds. Considering that existing RRSIS methods typically employ a simple and direct image-text alignment approach, neglecting the fine-grained relationships between images and text descriptions, FIANet [27] proposes fine-grained image-text alignment and text-aware multi-scale enhancement modules. These modules improve the segmentation of targets in complex remote sensing scenes.

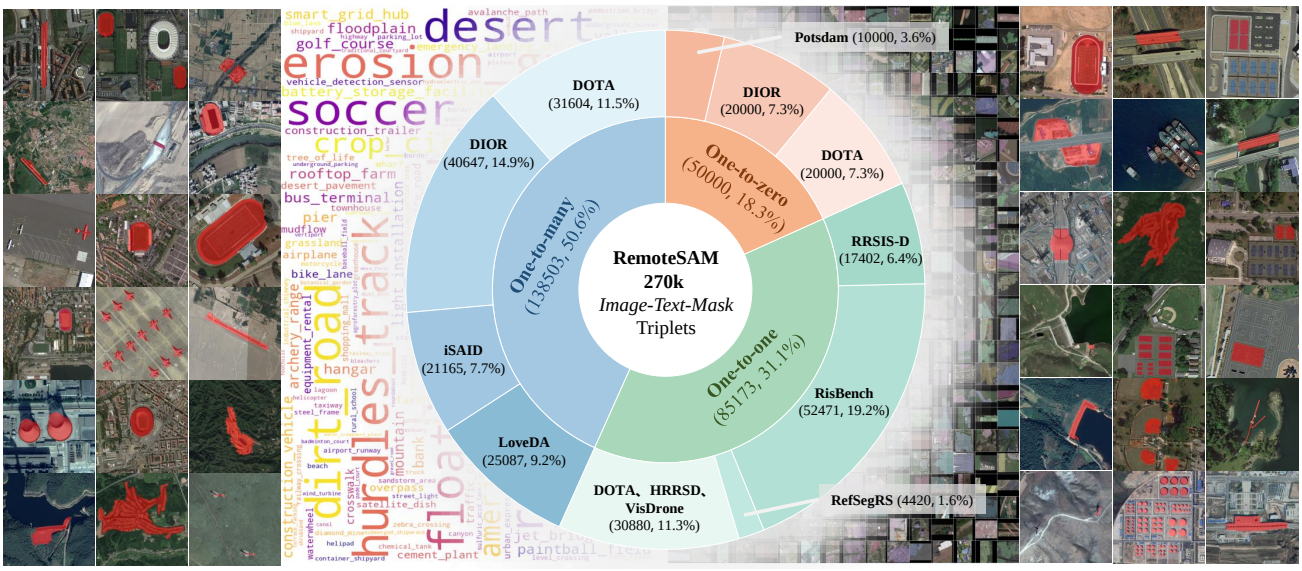


Figure 3. The composition of our proposed RemoteSAM-270K dataset. It is a generalized referring segmentation dataset, which is constructed through multi-source integration of mainstream remote sensing datasets, with semantically-dense referring expressions generated by VLMs.

3. RemoteSAM-270K Dataset

3.1. Improving Semantic Coverage of Data

Existing referring remote sensing image segmentation datasets exhibit limited semantic coverage, failing to support a foundation model's strong capabilities. To address this gap, we propose RemoteSAM-270K, a large-scale referring expression segmentation dataset that expands category coverage and attribute diversity for rich semantic coverage.

3.1.1. Category Expansion

As shown in Fig. 3, we propose to enrich the category completeness from existing remote sensing datasets. Specifically, we follow the steps outlined below to generate pixel-wise annotations.

Datasets Integration: We collect diverse remote sensing datasets (*e.g.*, iSAID [70], LoveDA [67], DOTA [72], HRRSD [84]) and standardize their formats. If the dataset is region-level, such as HRRSD [84], DOTA [72], we employ the code of SAMRS [65] to generate corresponding instance-level masks via their detection annotations.

Triplets Generation: We construct *image-text-mask* triplets for referring segmentation through three distinct data generation strategies, adhering to the generalized referring expression segmentation paradigm in benchmarks like G-RefCOCO [37] and RefZOM [19]: (1) "*One-to-One*": Directly integrate existing referring segmentation annotations from RefSegRS [79], RRSISD [43], and RisBench [17]. (2) "*One-to-Many*": Generate referring expressions following the template "{category} in the image." and decompose original segmentation masks into class-specific sub-masks. Each sub-mask aggregates all instances belonging to its corresponding category. (3) "*One-to-Zero*": Create null masks (all-zero matrices) paired with textual descriptions of categories explicitly absent from the image. This type of sample can effectively prevent the model from generating masks when the expressions refer to categories not present in the image.

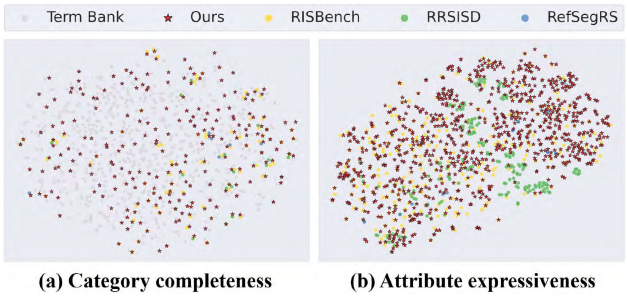
3.1.2. Attribute Expansion

Expressions with diverse attributes and flexible structures facilitate the model's capacity to learn rich semantic representations [6]. To achieve this objective, we integrate an automatic referring expression generation method with a semi-supervised Mixed-Teacher framework, producing high-quality *Image-Text-Mask* triplets enriched with detailed attributes.

Expressions Creation: Inspired by the Pyramid of Captions (PoCa) method [6], we split images into multiple local patches to expand the original remote sensing datasets (*e.g.*, DOTA [72], HRRSD [84]). Then, we employ Qwen2-VL-72B-Instruct [73] with the prompt "*Describe all of your observations in this image as comprehensively as possible.*" to generate attribute-enriched captions. However, these captions tend to focus on the information of the whole image rather than on a specific object. They cannot be directly utilized as referring expressions. Therefore, we design prompts (shown in Supplymentary) to further parse these captions into multiple descriptions, each containing only a single category target that has relevant referring information.

Table 1. Comparison of different referring remote sensing segmentation datasets. "Generalized": contains multi-target, no-target, and single-target expressions, "Cls": Categories, "Attr": Attributes.

Dataset	Generalized	# Samples	# Cls	# Attr	# Attr/Sample
RefSegRS [79]	×	4.4k	15	3	0.78
RRSIS-D [43]	×	17.4k	20	7	2.41
RISBench [17]	×	52.5k	26	8	2.45
RemoteSAM-270K	✓	**270K**	**297**	**16**	**3.17**



(a) Category completeness **(b) Attribute expressiveness**

Figure 4. Comparison of semantic coverage on ours and other remote sensing referring image segmentation datasets.

Masks Production: After obtaining the expressions, we also require a strategy for automatically generating masks. To achieve this objective, we utilize a mixture of several types of expert models (*e.g.*, GroundedSAM2 [55], RMSIN [43], etc.) to generate pseudo-labels. Unfortunately, even with multiple expert models, there are still a significant number of unreliable samples among the generated pseudo-labels. Therefore, to ensure data quality, we employ SigLIP2 [61] to calculate the similarity between the mask-related image regions and the corresponding expressions, removing samples with low similarities. This process is iterated, ultimately yields a set of attribute-rich triplets.

3.2. Data Analysis

We compare category and attribute distributions across four existing remote sensing referring image segmentation benchmarks. As illustrated in the Tab. 1, our RemoteSAM-270K dataset contains over 270,000 triplets of *image-text-mask*, spanning 297 categories and 16 types of fine-grained attribute descriptions. Furthermore, RemoteSAM-270K is the first generalized [19] remote sensing referring segmentation dataset, which can also enhance the depth of remote sensing referring segmentation research.

To facilitate the analysis of semantic coverage, we build a remote sensing vocabulary named RSBocab-1K. We integrate 1,000 fine-grained categories aligned with USGS Land Cover[1] and GB/T 21010-2017 remote sensing object classification specifications. Then, we organize these categories into three hierarchical levels. Following RSVocab-1K, we visualize the category completeness and attribute expressiveness utilizing t-SNE projection in Fig. 4. The visualization reveals two critical advantages of our RemoteSAM-270K: (1) superior category completeness and (2) richer attribute expressiveness. This enhanced comprehensiveness originates from our integration of cross-domain remote sensing datasets combined with VLMs that effectively mine latent semantic patterns.

[1]https://www.usgs.gov/programs/gap-analysis-project/science/land-cover

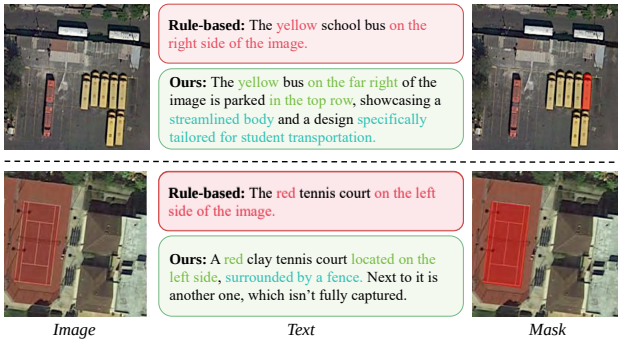


Figure 5. Comparison between VLMs (ours) and rule-based (RRSISD) generated expressions.

In Fig. 5, we showcase the advantages of referring expressions generated by the VLMs. It is evident that the rule-based generated expressions tend to have a more uniform structure, sometimes leading to ambiguities. For instance, the reference to "*the yellow school bus on the right side*" is not clear. In contrast, the expressions generated by VLMs include more detailed attribute descriptions and adhere to a more flexible syntax, which can help the model to learn more complex semantic information.

4. RemoteSAM

Pixel-level mask serves as the foundational computation unit in vision tasks, ensuring native compatibility with higher-level tasks at both region and image scales while preserving maximum spatial precision. Motivated by this insight, we build RemoteSAM through a unified task transition framework leveraging RES outputs through these strategies, as presented in Fig. 6.

Formally, given a training dataset $D_{\text{train}} = \{(I_k, T_k, M_k)\}_{k=1}^{N}$ composed of N *Image-Text-Mask* triplets, where each image $I_k \in \mathbb{R}^{H\times W\times 3}$ is paired with a referring expression $T_k = \{t_1, \ldots, t_n\}$ and its corresponding binary ground-truth mask $M_k \in \{0,1\}^{H\times W}$, RemoteSAM aims to predict a segmentation mask $\hat{M}$ for a query pair (I, T) at inference. The prediction is formulated as $\hat{M} = \mathcal{Q}(\mathcal{F}_v(I), \mathcal{F}_t(T))$, where $\mathcal{F}_v : \mathbb{R}^{H\times W\times 3} \to \mathbb{R}^{H'\times W'\times D}$ and $\mathcal{F}_t : \{w_i\}_{i=1}^{n} \to \mathbb{R}^{C}$ denote visual and textual encoders respectively, and $\mathcal{Q}$ is a fusion-decoder that jointly reasons over cross-modal

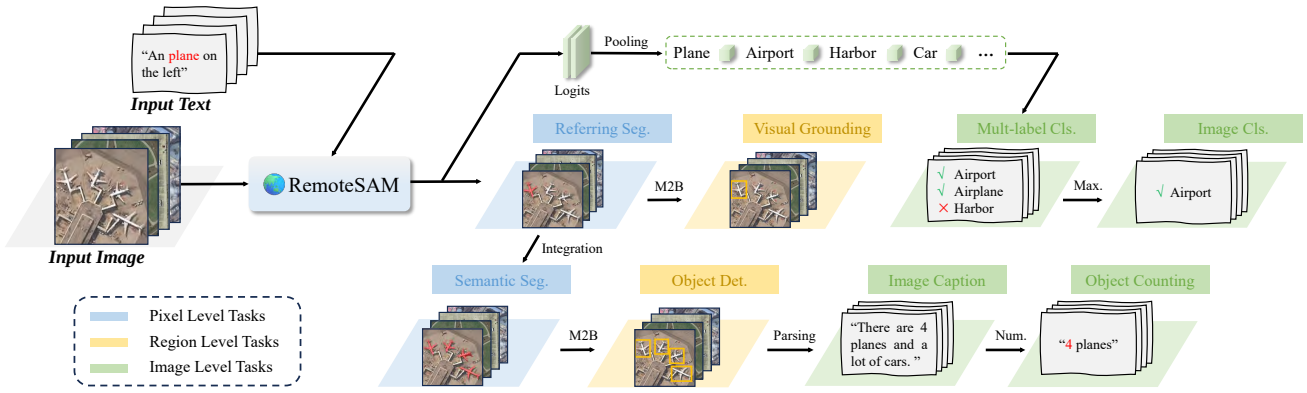


Figure 6. Overview of our proposed RemoteSAM. It is a foundational model centered around a referring expression segmentation framework. This generalist vision system demonstrates multi-task visual understanding capabilities through a single unified architecture without task-specific heads.

features to generate pixel-wise predictions. The model is trained by minimizing a segmentation loss $\mathcal{L}_{\text{seg}}$ between $\hat{M}$ and M.

Building upon the base prediction $\hat{M}$, we introduce a multi-task conversion strategy that transforms $\hat{M}$ into outputs for other vision tasks via task-specific functions $\{T_i : \hat{M} \rightarrow Y_i\}$, where i denotes the i^{th} task, Y_i defines the output space of task (*e.g.*, bounding boxes for detection, class probabilities for multi-label classification). Each task and its specific implementation steps are as follows.

4.1. Pixel-level Tasks

Referring Segmentation: Referring expression segmentation aims to segment specific objects or regions in an image via a free-form referring expression. This task serves as the foundational task for RemoteSAM and can directly utilize the model's original output mask $\hat{M}$.

Semantic Segmentation: Since we can obtain segmentation masks of any instance via generalized referring expressions, straightforwardly integrating all masks can produce a semantic segmentation result. Given a category set C for the specific segmentation task, for each class $c \in C$ we generate referring expression t_c = "All $\{c\}$ in the image" to acquire all the masks of this category:

$$\hat{M}_c = \mathcal{Q}\left(\mathcal{F}_v(I), \mathcal{F}_t(t_c)\right) \odot \left(P(c|I) \geq \tau_{\text{seg}}\right), \quad (1)$$

where $\hat{M}_c \in \{0,1\}^{H \times W}$ is the predicted mask, $P(c|I)$ is category confidence with τ_{seg} as threshold. Then, we can iteratively process all categories through the referring segmentation model, aggregating instance masks into semantic segmentation maps.

4.2. Region-level Tasks

Visual Grounding: Visual Grounding can be regarded as a type of "referring expression detection" task. Therefore, we directly convert predicted segmentation masks into bounding boxes via mask-to-bbox (M2B) [38] method $\mathcal{F}_{M2B}$. Specifically, for each predicted referring mask $\hat{M} \in \{0,1\}^{H \times W}$, compute grounding bounding box B coordinates:

$$B = \mathcal{F}_{M2B}(\hat{M}) = \left[(\min x_i,\ \min y_i), (\max x_j,\ \max y_j)\right]. \quad (2)$$

Object Detection: It shares conceptual similarities with semantic segmentation in visual recognition tasks. The primary distinction lies in their annotation formats: region-level bounding boxes and pixel-level masks. Therefore, semantic segmentation masks can be converted into rectangular bounding boxes. However, the M2B strategy encounters limitations when processing adjacent targets. Specifically, the overlapping regions of neighboring masks tend to merge during segmentation, resulting in a unified bounding box that encapsulates multiple distinct objects. To prevent adjacent objects from merging, we employ EPOC [7] to detect the objects' contours and refine mask boundaries. Then, we can utilize the M2B strategy (Eq. 2) to convert the semantic segmentation outputs into multiple bounding boxes.

4.3. Image-level Tasks

Multi-Label Classification: Since semantic segmentation maps indicate the existence of each category, intuitively, they can be converted into classification results. Specifically, we compute confidence scores through pooling over spatial-weighted probability maps of semantic segmentation outputs, followed by class-wise probability aggregation:

$$S_c = \lambda \cdot \frac{1}{HW} \sum_{i,j} P(c|i,j) + (1-\lambda) \cdot \max_{i,j} P(c|i,j), \quad (3)$$

where λ denotes balance parameters of max and average pooling, $P(c|i,j)$ represents the normalized probability of class c in position (i,j) of the input image.

Table 2. Comparison of semantic segmentation results on unseen dataset with various open-vocabulary and referring segmentation models.

Methods	Publication	Vaihingen	UDD5	DeepGlobe
		mIoU (%)	*mIoU* (%)	*mIoU* (%)
Open-vocabulary Models				
MaskCLIP [86]	ECCV22	24.7	32.4	13.2
SCLIP [66]	arXiv22	28.4	38.7	7.0
GEM [3]	CVPR24	24.7	41.2	4.7
ClearCLIP [26]	ECCV24	27.3	41.8	5.7
SegEarth-OV [29]	arXiv24	<u>29.1</u>	**50.6**	20.1
Referring Models				
FIANet [27]	TGRS24	1.1	2.4	47.8
RMSIN [43]	CVPR24	2.2	1.7	<u>47.9</u>
RemoteSAM	-	**46.0**	<u>45.6</u>	**60.5**

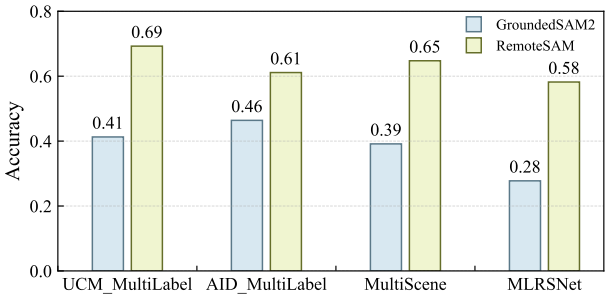


Figure 7. Comparison of zero-shot classification on SATIN with Grounded SAM2.

Then labels of class c are judged as positive when their corresponding confidence score S_c exceeds confidence threshold τ_{cls} (0.5 as default):

$$y_{multi} = \{S_k \geq \tau_{\text{cls}}\}. \tag{4}$$

Image Classification: Confidence scores are obtained in the same manner as multi-label classification, with the distinction being that we select the class with the highest confidence score:

$$y_{scene} = \arg\max\left[S_1, \ldots, S_C\right]. \tag{5}$$

Image Caption: When we possess object positions B, categories y_{multi}, and numbers N_c of objects, it is possible to generate captions in a rule-based manner [38, 80]. Specifically, y_{multi} provides an overview of image categories, while N_c and spatial relationships from B yield finer details.

Object Counting: As we can detect objects in the image, counting the number of class c_{target} can be easily performed:

$$N_c = \sum_{i=1}^{V} \mathbb{I}(c_i = c_{\text{target}}), \tag{6}$$

where $\mathbb{I}$ denotes the indicator function, V is the total number of detected bboxes, c_i is the i^{th} bbox's category.

5. Experiments

In this section, we conducted comprehensive experiments to evaluate the semantic coverage of RemoteSAM-270K and the task unification of RemoteSAM. Semantic coverage evaluates the generalization performance of the model trained on our dataset, including segmentation on unseen categories and zero-shot classification. Task unification evaluation is realized by testing the performance of various downstream visual-centric tasks, including pixel-level, region-level, and image-level. Additional experimental results are provided in the supplementary materials.

5.1. Evaluation Setup

5.1.1. Models

Our implementation built upon the RMSIN [43] architecture with BERT [16] as the textual encoder and Swin-Base [45] as the visual encoder. For downstream tasks, we compared three categories of candidates: (1) Vision pretrained foundation models (VFMs): RingMo [59], Scale-MAE [54], MA3E [33], et al; (2) Vision-language models (VLMs): GeoChat [25], LHRS-Bot [50], Falcon [76], et al; (3) Task-specific architectures: SegEarth-OV [29], GeoGround [87], MGVLF [80], et al.

5.1.2. Datasets

We compared our approach against previous SOTA referring segmentation models across two large-scale datasets: RRSISD [43] and RisBench [17]. The downstream evaluation protocol spans five established benchmarks in remote sensing: DOTA [72], DIOR [28], iSAID [70], Potsdam[2], and RSVG [80].

5.1.3. Training Settings.

The experimental configuration employed 8 × NVIDIA GeForce RTX 4090 GPUs with image resolution fixed at 896×896. Training proceeds for 40 epochs using AdamW optimization, initialized with a learning rate of 3e-5.

5.2. Evaluations of Semantic Coverage

In this section, we validated the semantic coverage of RemoteSAM-270K. We believe that the improvement in semantic coverage should lead to a qualitative enhancement in model performance, which can be demonstrated through the model's generalization abilities.

Firstly, we evaluated our RemoteSAM on unseen datasets, comparing it against several state-of-the-art open-vocabulary segmentation methods and other referring segmentation approaches, as illustrated in the Tab. 2. On the Vaihingen dataset, RemoteSAM outperformed SegEarth-OV by 16.9%, while other referring segmentation methods demonstrated significantly lower accuracy. These results

[2]https://www.isprs.org/education/benchmarks/UrbanSemLab/2d-sem-label-potsdam.aspx

Table 3. Comparison of referring expression segmentation results on RRSISD and RisBench.

Methods	Publication	RRSISD		RisBench	
		$oIoU(\%)$	$mIoU(\%)$	$oIoU(\%)$	$mIoU(\%)$
BRINet [20]	CVPR20	69.88	49.65	48.73	42.91
LSCM [23]	ECCV20	69.05	49.92	50.08	43.69
CMPC [22]	CVPR20	69.22	49.24	50.24	43.82
CMPC+ [42]	TPAMI21	70.13	50.12	53.98	46.73
LAVT [75]	CVPR22	77.19	61.04	74.15	61.93
CroosVLT [12]	TMM23	75.48	58.48	74.33	62.84
CARIS [44]	MM23	77.17	62.12	75.10	65.79
LGCE [79]	TGRS24	76.34	59.37	73.87	62.13
CroBIM [17]	TGRS24	75.99	64.46	73.04	68.03
robust-ref-seg [71]	TIP24	77.40	58.91	74.23	61.25
RMSIN [43]	CVPR24	77.88	64.26	75.24	68.25
RS2-SAM2 [57]	arxiv25	78.99	66.72	-	-
RemoteSAM	-	**80.04**	**71.75**	**75.93**	**71.46**

Table 4. Comparison of semantic segmentation results with various vision foundation models.

Methods	Publication	Pre-trained Data	iSAID	Potsdam
			$mIoU$ (%)	$mF1$ (%)
SeCo [48]	ICCV21	Sentinel-2 [52]	57.20	89.03
GASSL [74]	NIPS21	MillionAID [46]	65.95	91.27
SatMAE [14]	NIPS22	fMoW [13]	62.97	90.63
RingMo [59]	TGRS22	About 2M [59]	**67.20**	91.27
RVSA [64]	TGRS22	MillionAID [46]	64.49	-
SSL4EO [69]	GRSM23	SSL4EO-S12 [69]	64.01	91.54
CACo [47]	CVPR23	MillionAID [46]	64.32	91.35
SAMRS [65]	NIPS23	MillionAID [46]	66.26	91.43
ScaleMAE [54]	ICCV23	fMoW [13]	65.77	91.54
RSCoTr [31]	TGRS24	ImageNet-22k [15]	-	90.67
MA3E [33]	ECCV24	MillionAID [46]	64.06	91.50
RemoteSAM	-	RemoteSAM-270K	64.72	91.80
RemoteSAM-FT	-	RemoteSAM-270K	67.01	**93.54**

indicate that our dataset effectively enhances the model's cross-domain generalization competence alongside demonstrated open-set identification capacity.

To further validate the enhancement of category diversity in our dataset, we also conducted experiments on zero-shot classification utilizing the SATIN [56] meta dataset, as represented in Fig. 7. SATIN contains over 250 categories, comprised of images with various resolutions and viewpoints. The results indicate that the model trained on our RemoteSAM-270K outperforms the Grounded SAM2 [55] in recognition performance, which demonstrates that our data can effectively support the model in recognizing a majority of common remote sensing categories.

5.3. Evaluations of Task Unification

We further conducted a series of Downstream evaluations to measure the RemoteSAM's performance on three categories of downstream vision-centric tasks: pixel-level, region-level, and image-level tasks. Due to space limitations, we presented semantic segmentation, visual localization, multi-label classification, and object counting as the corresponding tasks for evaluation.

Table 5. Comparison of visual grounding performance with specialized and foundation models.

Methods	Publication	Parameters	RSVG	
			$AP_{50}(\%)$	$mIoU(\%)$
Specialized Models				
ZSGNet [58]	ICCV19	140M	51.67	44.12
FAOA [34]	CVPR20	150M	70.86	62.86
ResC [21]	CVPR21	150M	72.71	64.24
LBYL-Net [35]	TIP22	155M	73.78	65.92
TransVG [30]	NIPS21	136M	72.41	63.56
VLTVG [78]	CVPR22	155M	75.79	66.32
MGVLF [80]	TGRS23	136M	76.78	68.04
GeoGround [87]	arxiv24	7B	77.73	-
Foundation Models				
MiniGPT-v2 [88]	arXiv23	7B	46.64	-
Qwen-VL-Chat [73]	arXiv23	7B	44.76	-
SkyEyeGPT [81]	NIPS22	7B	70.50	-
LHRS-Bot [50]	ECCV24	7B	73.45	-
Falcon [76]	arXiv25	0.7B	56.87	-
RemoteSAM	-	**180M**	**74.36**	65.07

Table 6. Comparison of multi-label classification results with various remote sensing foundation models.

Methods	Publication	Parameters	DIOR	DOTAv2
			Acc	Acc
MiniCPM-V [18]	arXiv24	3B	0.08	0.10
MiniGPT-v2 [88]	arXiv23	7B	0.16	0.11
Qwen-VL-Chat [73]	arXiv23	7B	0.17	0.10
Sphinx [36]	arXiv23	7B	0.19	0.13
LLaVA-1.5 [41]	NIPS24	7B	0.35	0.19
RemoteCLIP [38]	TGRS24	390M	0.59	0.46
GeoChat [25]	CVPR24	7B	0.65	0.60
LHRS-Bot [50]	ECCV24	7B	0.35	0.20
Falcon [76]	arXiv25	7B	0.87	**0.95**
RemoteSAM	-	**180M**	**0.94**	**0.95**

Performance on Referring Expression Segmentation. As illustrated in Tab. 3, we evaluated referring segmentation performance on remote sensing imagery. The RemoteSAM demonstrates substantial superiority over comparative approaches. Specifically, it achieves 71.75% mIoU on RRSISD - surpassing the SAM2-based architecture (RS2-SAM2 [57]) by 5.03%. Our approach also establishes a new state-of-the-art result on RisBench with 3.21% absolute performance gain. This advancement primarily stems from the extensive semantic coverage in our training dataset.

Performance on Semantic Segmentation. To validate the performance of RemoteSAM on the semantic segmentation task, we selected several vision pre-trained models for comparison. As presented in Tab. 4, the results indicate that our approach performs comparably to specialized vision backbone models without task-specific tuning and even attains state-of-the-art accuracy on the Potsdam dataset (91.80%). Moreover, further performance improvements (67.01% and 93.54%) can be achieved by training the model on the specific dataset.

Performance on Visual Grounding. Beyond pixel-level tasks, our RemoteSAM also supports fine-grained

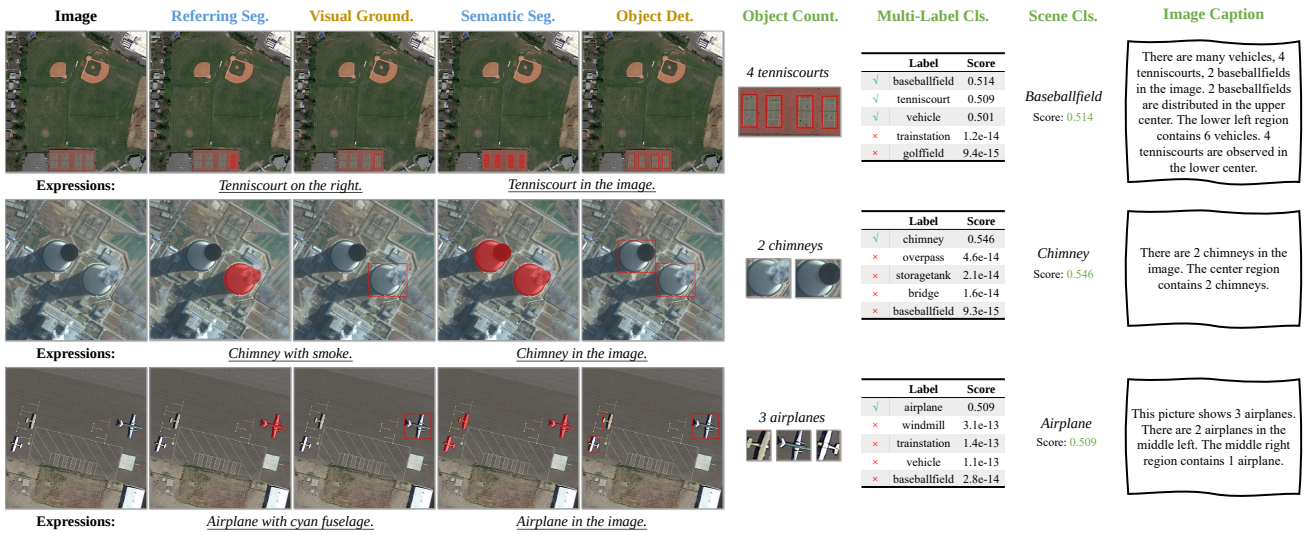


Figure 8. Inference examples of RemoteSAM on 8 visual-centric tasks.

Table 7. Comparison of object counting results with various remote sensing foundation models.

Methods	Publication	Parameters	DIOR	DOTAv2
			Acc	*Acc*
MiniCPM-V [18]	arXiv24	3B	0.426	0.260
MiniGPT-v2 [88]	arXiv23	7B	0.429	0.248
Sphinx [36]	arXiv23	7B	0.430	0.257
LLaVA-1.5 [41]	NIPS24	7B	0.249	0.221
GeoChat [25]	CVPR24	7B	0.453	0.240
LHRS-Bot [50]	ECCV24	7B	0.455	0.244
RemoteSAM	-	**180M**	**0.620**	**0.409**

region-level tasks. For example, we present the visual grounding performance of RemoteSAM on the RSVG [80] dataset, comparing it against specialized and foundation models. As illustrated in Tab. 5, RemoteSAM performed as accurately as specialized models. Compared to other VLMs, our approach outperformed them significantly while utilizing a substantially smaller number of parameters.

Performance on Multi-label Classification. We evaluated the performance of RemoteSAM and VLMs on multi-label classification, with the results presented in Tab. 6. In the multi-label classification task, RemoteSAM achieved accuracy rates of 94% and 95% on the DIOR and DOTA datasets, surpassing GeoChat by 29% and 35%, respectively. This improvement may stem from the model's remarkable capability to comprehend complex spatial relationships in remote sensing scenes compared to VLMs.

Performance on Object Counting. To further validate the performance of RemoteSAM at the image level, we conducted experiments on the object counting task. The task requires the model's ability to perceive and reason compositionally. As presented in Tab. 7, our RemoteSAM attained accuracy rates of 62.0% and 40.9%, significantly outperforming other approaches such as LHRS-Bot.

The above results demonstrate that our referring segmentation-centered visual unification paradigm achieves remarkable performance across pixel-level, region-level, and image-level tasks. It highlights RemoteSAM's ability to handle complex remote sensing tasks. More importantly, RemoteSAM has a significantly smaller number of parameters compared to other large-scale vision-language models and does not require task-specific decoders, enabling it to efficiently process high-resolution remote sensing images.

5.4. Further Analysis

We presented a qualitative analysis of RemoteSAM across eight visual tasks in Fig. 8. Overall, RemoteSAM successfully translates the results of referring expression segmentation into outputs required for various other visual-centric tasks with precision, ranging from pixel-level to image-level. From a variety of examples, it is evident that RemoteSAM is capable of (1) localizing the boundaries of specified targets accurately (referring segmentation, semantic segmentation); (2) analyzing the relationships between objects in images (object detection, visual grounding); and (3) understanding the global contextual information in images (counting, classification, captioning). Specifically, in addition to locating objects via attributes such as location (first row), RemoteSAM also comprehends status information. For example, in the second row, it identifies a chimney that is smoking. Moreover, when provided with a generalized 'one-to-many' expression (*e.g.*, "Airplane in the image."), RemoteSAM can further execute additional visual tasks.

6. Conclusion

In this work, we present **RemoteSAM**, a unified visual foundation model for Earth observation that addresses the critical pixel-level limitations of existing remote sensing foundation models through a referring segmentation-based paradigm. By developing an automated data curation pipeline leveraging VLMs and multi-teacher localization, we construct the largest referring expression segmentation dataset, RemoteSAM-270K (270K *image-text-mask* triplets), with significant semantic diversity spanning 297 categories and 16 attributes. We also build a hierarchical semantic vocabulary to evaluate the semantic coverage of remote sensing datasets. Extensive evaluations demonstrate RemoteSAM's superiority in handling classification, detection, segmentation, and grounding with significant parameter efficiency. Our work demonstrates that segmentation-centric architectures can serve as unified backbones for multimodal Earth observation intelligence.

Acknowledge

This work was supported in part by the National Natural Science Foundation of China under Grant 62372155 and Grant 62302149, in part by the Aeronautical Science Fund under Grant 2022Z071108001, in part by the Qinglan Project of Jiangsu Province, and in part by Changzhou Science and Technology Bureau Project No. 20231313

References

[1] Edoardo Arnaudo, Jacopo Lungo Vaschetti, Lorenzo Innocenti, Luca Barco, Davide Lisi, Vanina Fissore, and Claudio Rossi. Fmars: Annotating remote sensing images for disaster management using foundation models. In *IEEE IGARSS*, pages 3920–3924. IEEE, 2024. 1

[2] Muhammad Awais, Muzammal Naseer, Salman Khan, Rao Muhammad Anwer, Hisham Cholakkal, Mubarak Shah, Ming-Hsuan Yang, and Fahad Shahbaz Khan. Foundation models defining a new era in vision: a survey and outlook. *IEEE TPAMI*, 2025. 1, 3

[3] Walid Bousselham, Felix Petersen, Vittorio Ferrari, and Hilde Kuehne. Grounding everything: Emerging localization properties in vision-language transformers. In *CVPR*, page 3828–3837. IEEE, 2024. 7

[4] Christopher Brown, Michal Kazmierski, William Rucklidge, Valerie Pasquarella, and Evan Shelhamer. Learned embedding fields for multi-source, multi-temporal earth observation imagery. In *ICLR Workshop on Machine Learning for Remote Sensing (ML4RS)*, 2024. 1

[5] Keumgang Cha, Junghoon Seo, and Taekyung Lee. A billion-scale foundation model for remote sensing images. *IEEE JSTARS*, page 1–17, 2024. 3

[6] Delong Chen, Samuel Cahyawijaya, Etsuko Ishii, Ho Shu Chan, Yejin Bang, and Pascale Fung. What makes for good image captions?, 2024. 4

[7] Delong Chen, Samuel Cahyawijaya, Jianfeng Liu, Baoyuan Wang, and Pascale Fung. Subobject-level image tokenization. *arXiv preprint arXiv:2402.14327*, 2024. 6

[8] Gang Chen, Yuyu Zhou, James A Voogt, and Eleanor C Stokes. Remote sensing of diverse urban environments: From the single city to multiple cities. *RSE*, 305:114108, 2024. 1

[9] Ting Chen, Saurabh Saxena, Lala Li, David J. Fleet, and Geoffrey Hinton. Pix2seq: A language modeling framework for object detection, 2022. 2

[10] Ting Chen, Saurabh Saxena, Lala Li, Tsung-Yi Lin, David J. Fleet, and Geoffrey Hinton. A unified sequence interface for vision tasks, 2022. 2

[11] Yong Xien Chng, Henry Zheng, Yizeng Han, Xuchong Qiu, and Gao Huang. Mask grounding for referring image segmentation. In *CVPR*, pages 26573–26583, 2024. 2, 3

[12] Yubin Cho, Hyunwoo Yu, and Suk-Ju Kang. Cross-aware early fusion with stage-divided vision and language transformer encoders for referring image segmentation. *IEEE TMM*, 26:5823–5833, 2023. 8

[13] Gordon Christie, Neil Fendley, James Wilson, and Ryan Mukherjee. Functional map of the world. In *CVPR*, pages 6172–6180, 2018. 8

[14] Yezhen Cong, Samar Khanna, Chenlin Meng, Patrick Liu, Erik Rozi, Yutong He, Marshall Burke, David Lobell, and Stefano Ermon. Satmae: Pre-training transformers for temporal and multi-spectral satellite imagery. *NeurIPS*, 35:197–211, 2022. 2, 3, 8

[15] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, pages 248–255. Ieee, 2009. 8

[16] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT*, pages 4171–4186, 2019. 7

[17] Zhe Dong, Yuzhe Sun, Yanfeng Gu, and Tianzhu Liu. Cross-modal bidirectional interaction model for referring remote sensing image segmentation. *arXiv preprint arXiv:2410.08613*, 2024. 1, 3, 4, 5, 7, 8

[18] Shengding Hu, Yuge Tu, Xu Han, Chaoqun He, Ganqu Cui, Xiang Long, Zhi Zheng, Yewei Fang, Yuxiang Huang, Weilin Zhao, et al. Minicpm: Unveiling the potential of small language models with scalable training strategies. *arXiv preprint arXiv:2404.06395*, 2024. 8, 9

[19] Yutao Hu, Qixiong Wang, Wenqi Shao, Enze Xie, Zhenguo Li, Jungong Han, and Ping Luo. Beyond one-to-one: Rethinking the referring image segmentation. In *ICCV*, pages 4067–4077, 2023. 2, 4, 5

[20] Zhiwei Hu, Guang Feng, Jiayu Sun, Lihe Zhang, and Huchuan Lu. Bi-directional relationship inferring network for referring image segmentation. In *CVPR*, pages 4424–4433, 2020. 8

[21] Binbin Huang, Dongze Lian, Weixin Luo, and Shenghua Gao. Look before you leap: Learning landmark features for one-stage visual grounding. In *CVPR*, pages 16888–16897, 2021. 8

[22] Shaofei Huang, Tianrui Hui, Si Liu, Guanbin Li, Yunchao Wei, Jizhong Han, Luoqi Liu, and Bo Li. Referring image segmentation via cross-modal progressive comprehension. In *CVPR*, pages 10488–10497, 2020. 8

[23] Tianrui Hui, Si Liu, Shaofei Huang, Guanbin Li, Sansi Yu, Faxi Zhang, and Jizhong Han. Linguistic structure guided context modeling for referring image segmentation. In *ECCV*, pages 59–75. Springer, 2020. 8

[24] Lixia Ji, Yunlong Du, Yiping Dang, Wenzhao Gao, and Han Zhang. A survey of methods for addressing the challenges of referring image segmentation. *Neurocomputing*, 583:127599, 2024. 2, 3

[25] 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 *CVPR*, pages 27831–27840, 2024. 1, 2, 3, 7, 8, 9

[26] Mengcheng Lan, Chaofeng Chen, Yiping Ke, Xinjiang Wang, Litong Feng, and Wayne Zhang. Clearclip: Decomposing clip representations for dense vision-language inference. In *ECCV*, pages 143–160. Springer, 2024. 7

[27] Sen Lei, Xinyu Xiao, Tianlin Zhang, Heng-Chao Li, Zhenwei Shi, and Qing Zhu. Exploring fine-grained image-text alignment for referring remote sensing image segmentation. *IEEE TGRS*, 2024. 3, 7

[28] Ke Li, Gang Wan, Gong Cheng, Liqiu Meng, and Junwei Han. Object detection in optical remote sensing images: A survey and a new benchmark. *ISPRS PRS*, 159:296–307, 2020. 7

[29] Kaiyu Li, ruixun Liu, Xiangyong Cao, Xueru Bai, Feng Zhou, Deyu Meng, and Zhi Wang. Segearth-ov: Towards training-free open-vocabulary segmentation for remote sensing images. *arXiv preprint arXiv:2410.01768*, 2024. 7

[30] Muchen Li and Leonid Sigal. Referring transformer: A one-step approach to multi-task visual grounding. *NeurIPS*, 34: 19652–19664, 2021. 8

[31] Qingyun Li, Yushi Chen, Xin He, and Lingbo Huang. Co-training transformer for remote sensing image classification, segmentation and detection. *IEEE TGRS*, 62:1–18, 2024. 8

[32] Yan Li, Weiwei Guo, Xue Yang, Ning Liao, Dunyun He, Jiaqi Zhou, and Wenxian Yu. Toward open vocabulary aerial object detection with clip-activated student-teacher learning. In *ECCV*, pages 431–448. Springer, 2024. 1

[33] Zhihao Li, Biao Hou, Siteng Ma, Zitong Wu, Xianpeng Guo, Bo Ren, and Licheng Jiao. Masked angle-aware autoencoder for remote sensing images. In *ECCV*, pages 260–278. Springer, 2024. 2, 3, 7, 8

[34] Yue Liao, Si Liu, Guanbin Li, Fei Wang, Yanjie Chen, Chen Qian, and Bo Li. A real-time cross-modality correlation filtering method for referring expression comprehension. In *CVPR*, pages 10880–10889, 2020. 8

[35] Yue Liao, Aixi Zhang, Zhiyuan Chen, Tianrui Hui, and Si Liu. Progressive language-customized visual feature learning for one-stage visual grounding. *IEEE TIP*, 31:4266–4277, 2022. 8

[36] Ziyi Lin, Chris Liu, Renrui Zhang, Peng Gao, Longtian Qiu, Han Xiao, Han Qiu, Chen Lin, Wenqi Shao, Keqin Chen, et al. Sphinx: The joint mixing of weights, tasks, and visual embeddings for multi-modal large language models. *arXiv preprint arXiv:2311.07575*, 2023. 8, 9

[37] Chang Liu, Henghui Ding, and Xudong Jiang. Gres: Generalized referring expression segmentation. In *CVPR*, pages 23592–23601, 2023. 4

[38] Fan Liu, Delong Chen, Zhangqingyun Guan, Xiaocong Zhou, Jiale Zhu, Qiaolin Ye, Liyong Fu, and Jun Zhou. Remoteclip: A vision language foundation model for remote sensing. *IEEE TGRS*, 2024. 1, 6, 7, 8

[39] Fan Liu, Liang Yao, Chuanyi Zhang, Ting Wu, Xinlei Zhang, Xiruo Jiang, and Jun Zhou. Scale-invariant feature disentanglement via adversarial learning for uav-based object detection. *arXiv preprint arXiv:2405.15465*, 2024. 3

[40] Fan Liu, Tianshu Zhang, Wenwen Dai, Chuanyi Zhang, Wenwen Cai, Xiaocong Zhou, and Delong Chen. Few-shot adaptation of multi-modal foundation models: A survey. *Artificial Intelligence Review*, 57(10):268, 2024. 3

[41] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning, 2023. 8, 9

[42] Si Liu, Tianrui Hui, Shaofei Huang, Yunchao Wei, Bo Li, and Guanbin Li. Cross-modal progressive comprehension for referring segmentation. *IEEE TPAMI*, 44(9):4761–4775, 2021. 8

[43] Sihan Liu, Yiwei Ma, Xiaoqing Zhang, Haowei Wang, Jiayi Ji, Xiaoshuai Sun, and Rongrong Ji. Rotated multi-scale interaction network for referring remote sensing image segmentation. In *CVPR*, pages 26658–26668, 2024. 2, 3, 4, 5, 7, 8

[44] Sun-Ao Liu, Yiheng Zhang, Zhaofan Qiu, Hongtao Xie, Yongdong Zhang, and Ting Yao. Caris: Context-aware referring image segmentation. In *ACM MM*, pages 779–788, 2023. 8

[45] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *ICCV*, pages 10012–10022, 2021. 7

[46] Yang Long, Gui-Song Xia, Shengyang Li, Wen Yang, Michael Ying Yang, Xiao Xiang Zhu, Liangpei Zhang, and Deren Li. On creating benchmark dataset for aerial image interpretation: Reviews, guidances, and million-aid. *IEEE JSTARS*, 14:4205–4230, 2021. 8

[47] Utkarsh Mall, Bharath Hariharan, and Kavita Bala. Change-aware sampling and contrastive learning for satellite images. In *CVPR*, pages 5261–5270, 2023. 8

[48] Oscar Manas, Alexandre Lacoste, Xavier Giró-i Nieto, David Vazquez, and Pau Rodriguez. Seasonal contrast: Unsupervised pre-training from uncurated remote sensing data. In *ICCV*, pages 9414–9423, 2021. 8

[49] Shiyu Miao, Delong Chen, Fan Liu, Chuanyi Zhang, Yanhui Gu, Shengjie Guo, and Jun Zhou. Prompting directsam for semantic contour extraction in remote sensing images. In *ICASSP*, pages 1–5. IEEE, 2025. 3

[50] Dilxat Muhtar, Zhenshi Li, Feng Gu, Xueliang Zhang, and Pengfeng Xiao. Lhrs-bot: Empowering remote sensing with vgi-enhanced large multimodal language model. In *ECCV*, pages 440–457. Springer, 2024. 1, 2, 3, 7, 8, 9

[51] Mubashir Noman, Muzammal Naseer, Hisham Cholakkal, Rao Muhammad Anwar, Salman Khan, and Fahad Shahbaz Khan. Rethinking transformers pre-training for multi-spectral satellite imagery. In *CVPR*, pages 27811–27819, 2024. 3

[52] Darius Phiri, Matamyo Simwanda, Serajis Salekin, Vincent R Nyirenda, Yuji Murayama, and Manjula Ranagalage. Sentinel-2 data for land cover/use mapping: A review. *Remote sensing*, 12(14):2291, 2020. 8

[53] Colorado J Reed, Ritwik Gupta, Shufan Li, Sarah Brockman, Christopher Funk, Brian Clipp, Kurt Keutzer, Salvatore Candido, Matt Uyttendaele, and Trevor Darrell. Scale-mae: A scale-aware masked autoencoder for multiscale geospatial representation learning. In *ICCV*, pages 4065–4076, 2023. 1, 3

[54] Colorado J Reed, Ritwik Gupta, Shufan Li, Sarah Brockman, Christopher Funk, Brian Clipp, Kurt Keutzer, Salvatore Candido, Matt Uyttendaele, and Trevor Darrell. Scale-mae: A scale-aware masked autoencoder for multiscale geospatial representation learning. In *ICCV*, pages 4088–4099, 2023. 3, 7, 8

[55] Tianhe Ren, Shilong Liu, Ailing Zeng, Jing Lin, Kunchang Li, He Cao, Jiayu Chen, Xinyu Huang, Yukang Chen, Feng Yan, Zhaoyang Zeng, Hao Zhang, Feng Li, Jie Yang, Hongyang Li, Qing Jiang, and Lei Zhang. Grounded sam: Assembling open-world models for diverse visual tasks, 2024. 5, 8

[56] Jonathan Roberts, Kai Han, and Samuel Albanie. Satin: A multi-task metadataset for classifying satellite imagery using vision-language models. *arXiv preprint arXiv:2304.11619*, 2023. 8

[57] Fu Rong, Meng Lan, Qian Zhang, and Lefei Zhang. Customized sam 2 for referring remote sensing image segmentation, 2025. 8

[58] Arka Sadhu, Kan Chen, and Ram Nevatia. Zero-shot grounding of objects from natural language queries. In *ICCV*, pages 4694–4703, 2019. 8

[59] Xian Sun, Peijin Wang, Wanxuan Lu, Zicong Zhu, Xiaonan Lu, Qibin He, Junxi Li, Xuee Rong, Zhujun Yang, Hao Chang, et al. Ringmo: A remote sensing foundation model with masked image modeling. *IEEE TGRS*, 61:1–22, 2022. 1, 7, 8

[60] Xian Sun, Peijin Wang, Wanxuan Lu, Zicong Zhu, Xiaonan Lu, Qibin He, Junxi Li, Xuee Rong, Zhujun Yang, Hao Chang, Qinglin He, Guang Yang, Ruiping Wang, Jiwen Lu, and Kun Fu. Ringmo: A remote sensing foundation model with masked image modeling. *IEEE TGRS*, 61:1–22, 2023. 3

[61] Michael Tschannen, Alexey Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim Alabdulmohsin, Nikhil Parthasarathy, Talfan Evans, Lucas Beyer, Ye Xia, Basil Mustafa, Olivier Hénaff, Jeremiah Harmsen, Andreas Steiner, and Xiaohua Zhai. Siglip 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features. *arXiv preprint arXiv:2502.14786*, 2025. 5

[62] Devis Tuia, Konrad Schindler, Begüm Demir, Xiao Xiang Zhu, Mrinalini Kochupillai, Sašo Džeroski, Jan N van Rijn, Holger H Hoos, Fabio Del Frate, Mihai Datcu, et al. Artificial intelligence to advance earth observation: A review of models, recent trends, and pathways forward. *IEEE GRSM*, 2024. 1

[63] Nancy Victor, Praveen Kumar Reddy Maddikunta, Delphin Raj Kesari Mary, Ramalingam Murugan, Rajeswari Chengoden, Thippa Reddy Gadekallu, Nitin Rakesh, Yaodong Zhu, and Jeongyeup Paek. Remote sensing for agriculture in the era of industry 5.0–a survey. *IEEE JSTARS*, 2024. 1

[64] Di Wang, Qiming Zhang, Yufei Xu, Jing Zhang, Bo Du, Dacheng Tao, and Liangpei Zhang. Advancing plain vision transformer toward remote sensing foundation model. *IEEE TGRS*, 61:1–15, 2022. 8

[65] Di Wang, Jing Zhang, Bo Du, Minqiang Xu, Lin Liu, Dacheng Tao, and Liangpei Zhang. Samrs: Scaling-up remote sensing segmentation dataset with segment anything model. *NeurIPS*, 36:8815–8827, 2023. 4, 8

[66] Feng Wang, Jieru Mei, and Alan Yuille. Sclip: Rethinking self-attention for dense vision-language inference. *arXiv preprint arXiv:2312.01597*, 2023. 7

[67] Junjue Wang, Zhuo Zheng, Ailong Ma, Xiaoyan Lu, and Yanfei Zhong. Loveda: A remote sensing land-cover dataset for domain adaptive semantic segmentation. *arXiv preprint arXiv:2110.08733*, 2021. 4

[68] Peng Wang, An Yang, Rui Men, Junyang Lin, Shuai Bai, Zhikang Li, Jianxin Ma, Chang Zhou, Jingren Zhou, and Hongxia Yang. Ofa: Unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework. In *ICML*, pages 23318–23340. PMLR, 2022. 2

[69] Yi Wang, Nassim Ait Ali Braham, Zhitong Xiong, Chenying Liu, Conrad M Albrecht, and Xiao Xiang Zhu. Ssl4eo-s12: A large-scale multimodal, multitemporal dataset for self-supervised learning in earth observation [software and data sets]. *IEEE GRSM*, 11(3):98–106, 2023. 8

[70] Syed Waqas Zamir, Aditya Arora, Akshita Gupta, Salman Khan, Guolei Sun, Fahad Shahbaz Khan, Fan Zhu, Ling Shao, Gui-Song Xia, and Xiang Bai. isaid: A large-scale dataset for instance segmentation in aerial images. In *CVPRW*, pages 28–37, 2019. 4, 7

[71] Jianzong Wu, Xiangtai Li, Xia Li, Henghui Ding, Yunhai Tong, and Dacheng Tao. Towards robust referring image segmentation. *IEEE TIP*, 2024. 2, 8

[72] 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 *CVPR*, pages 3974–3983, 2018. 4, 7

[73] An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, and Chengyuan Li et al. Qwen2.5 technical report, 2025. 2, 3, 4, 8

[74] Longqi Yang, Liangliang Zhang, and Wenjing Yang. Graph adversarial self-supervised learning. *NeurIPS*, 34:14887–14899, 2021. 8

[75] Zhao Yang, Jiaqi Wang, Yansong Tang, Kai Chen, Hengshuang Zhao, and Philip HS Torr. Lavt: Language-aware vision transformer for referring image segmentation. In *CVPR*, pages 18155–18165, 2022. 8

[76] Kelu Yao, Nuo Xu, Rong Yang, Yingying Xu, Zhuoyan Gao, Titinunt Kitrungrotsakul, Yi Ren, Pu Zhang, Jin Wang, Ning

Wei, et al. Falcon: A remote sensing vision-language foundation model. *arXiv preprint arXiv:2503.11070*, 2025. 1, 2, 7, 8

[77] Liang Yao, Fan Liu, Chuanyi Zhang, Zhiquan Ou, and Ting Wu. Domain-invariant progressive knowledge distillation for uav-based object detection. *IEEE GRSL*, 2024. 3

[78] Jiabo Ye, Junfeng Tian, Ming Yan, Xiaoshan Yang, Xuwu Wang, Ji Zhang, Liang He, and Xin Lin. Shifting more attention to visual backbone: Query-modulated refinement networks for end-to-end visual grounding. In *CVPR*, pages 15502–15512, 2022. 8

[79] Zhenghang Yuan, Lichao Mou, Yuansheng Hua, and Xiao Xiang Zhu. Rrsis: Referring remote sensing image segmentation. *IEEE TGRS*, 2024. 2, 3, 4, 5, 8

[80] Yang Zhan, Zhitong Xiong, and Yuan Yuan. Rsvg: Exploring data and models for visual grounding on remote sensing data. *IEEE TGRS*, 61:1–13, 2023. 7, 8, 9

[81] Yang Zhan, Zhitong Xiong, and Yuan Yuan. Skyeyegpt: Unifying remote sensing vision-language tasks via instruction tuning with large language model. *ISPRS PRS*, 221:64–77, 2025. 8

[82] Jingyi Zhang, Jiaxing Huang, Sheng Jin, and Shijian Lu. Vision-language models for vision tasks: A survey. *IEEE TPAMI*, 2024. 1

[83] Wei Zhang, Miaoxin Cai, Tong Zhang, Yin Zhuang, and Xuerui Mao. Earthgpt: A universal multi-modal large language model for multi-sensor image comprehension in remote sensing domain. *IEEE TGRS*, 2024. 2

[84] Yuanlin Zhang, Yuan Yuan, Yachuang Feng, and Xiaoqiang Lu. Hierarchical and robust convolutional neural network for very high-resolution remote sensing object detection. *IEEE TGRS*, 57(8):5535–5548, 2019. 4

[85] Zilun Zhang, Tiancheng Zhao, Yulong Guo, and Jianwei Yin. Rs5m and georsclip: A large scale vision-language dataset and a large vision-language model for remote sensing. *IEEE TGRS*, 2024. 1

[86] Chong Zhou, Chen Change Loy, and Bo Dai. Extract free dense labels from clip. In *ECCV*, 2022. 7

[87] Yue Zhou, Mengcheng Lan, Xiang Li, Yiping Ke, Xue Jiang, Litong Feng, and Wayne Zhang. Geoground: A unified large vision-language model. for remote sensing visual grounding. *arXiv preprint arXiv:2411.11904*, 2024. 7, 8

[88] Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*, 2023. 3, 8, 9

Appendix

A. Quantitative comparison results for remaining tasks

In this section, we present the performance of RemoteSAM for remaining tasks, including Referring Segmentation (c.f. Tab. 8-Tab. 10), Image Caption (c.f. Tab. 11) and Object Detection (c.f. Tab. 12-Tab. 13). Detailed comparisons of Complex Scenes Classification task in SATIN meta dataset are also depicted (c.f. Fig. 9-Fig. 13).

As shown in Tables 8 to 10, RemoteSAM demonstrates exceptional performance on Referring Segmentation tasks, achieving top results across nearly all evaluated metrics on the three datasets, with only minor deviations in a few cases. For Image Caption task in Tab. 11, our rule-based captioning strategy attains a competitive CIDEr score of 12.370 on the UCM-Captions dataset compared to generic foundation models. Furthermore, RemoteSAM also excels in Object Detection tasks, achieving AP_{50} scores of 62.74% and 34.41% on the DIOR and iSAID datasets, respectively, surpassing other foundation models. Besides, we also present category-level comparisons between RemoteSAM and GroundedSAM2, which strongly evidences that scaled semantic coverage does translate to improved generalization capability.

Table 8. Detailed comparison of Referring Segmentation performance on RisBench.

Methods	Publication	RisBench						
		$Pr@0.5$	$Pr@0.6$	$Pr@0.7$	$Pr@0.8$	$Pr@0.9$	$oIoU(\%)$	$mIoU(\%)$
BRINet	CVPR20	52.87	45.39	38.64	30.79	11.86	48.73	42.91
LSCM	ECCV20	55.26	47.14	40.10	33.29	13.91	50.08	43.69
CMPC	CVPR20	55.17	47.84	40.28	32.87	14.55	50.24	43.82
CMPC+	TPAMI21	58.02	49.00	42.53	35.26	17.88	53.98	46.73
LAVT	CVPR22	69.40	63.66	56.10	44.95	25.21	74.15	61.93
CroosVLT	TMM23	70.62	65.05	57.40	45.80	26.10	74.33	62.84
CARIS	MM23	73.94	68.93	62.08	50.31	29.08	75.10	65.79
LGCE	TGRS24	69.64	64.07	56.26	44.92	25.74	73.87	62.13
CroBIM	TGRS24	77.55	72.83	66.38	55.93	34.07	73.04	69.33
robust-ref-seg	TIP24	69.15	63.24	55.33	43.27	24.20	74.23	61.25
RMSIN	CVPR24	76.32	71.53	64.83	54.2	33.76	75.24	68.25
RemoteSAM	-	**79.16**	**74.24**	**67.74**	**58.09**	**38.80**	**75.93**	**71.46**

Table 9. Detailed comparison of Referring Segmentation performance on RRSISD.

Methods	Publication	RRSISD						
		$Pr@0.5$	$Pr@0.6$	$Pr@0.7$	$Pr@0.8$	$Pr@0.9$	$oIoU(\%)$	$mIoU(\%)$
BRINet	CVPR20	56.90	48.77	39.12	27.03	8.73	69.88	49.65
LSCM	ECCV20	56.02	46.25	37.7	25.28	8.27	69.05	49.92
CMPC	CVPR20	55.83	47.40	36.94	25.45	9.19	69.22	49.24
CMPC+	TPAMI21	57.95	48.31	37.61	24.33	7.94	70.13	50.12
LAVT	CVPR22	69.52	63.63	53.29	42.55	24.53	77.19	61.04
CroosVLT	TMM23	66.42	59.41	49.76	38.67	23.30	75.48	58.48
CARIS	MM23	71.50	63.52	52.92	40.94	23.90	77.17	62.12
LGCE	TGRS24	67.65	61.53	51.45	39.62	23.33	76.34	59.37
CroBIM	TGRS24	74.58	67.57	55.59	41.63	23.56	75.99	64.46
robust-ref-seg	TIP24	66.59	59.58	49.93	38.72	23.30	77.40	58.91
RMSIN	CVPR24	74.69	68.40	56.54	42.95	24.76	77.88	64.26
RS2-SAM2	arXiv25	77.56	72.34	61.76	47.92	**29.73**	78.99	66.72
RemoteSAM	-	**84.46**	**78.45**	**66.25**	**49.32**	29.62	**80.04**	**71.75**

Table 10. Detailed comparison of Referring Segmentation performance on RefSegRS.

Methods	Publication	RefSegRS						
		$Pr@0.5$	$Pr@0.6$	$Pr@0.7$	$Pr@0.8$	$Pr@0.9$	$oIoU(\%)$	$mIoU(\%)$
BRINet	CVPR20	22.56	15.74	9.85	3.52	0.5	60.16	32.87
LAVT	CVPR22	71.44	57.40	32.14	15.41	4.51	<u>76.46</u>	57.74
LGCE	TGRS24	73.75	61.14	<u>39.46</u>	<u>16.02</u>	**5.45**	**76.81**	<u>59.96</u>
CroBIM	TGRS24	<u>75.89</u>	<u>61.42</u>	34.07	12.99	2.75	72.33	59.77
RMSIN	CVPR24	68.63	52.61	26.47	10.13	1.82	71.46	55.71
RemoteSAM	-	**79.69**	**70.89**	**54.93**	**24.11**	<u>5.01</u>	75.49	**65.79**

Table 11. Comparison of zero-shot Image Caption performance with Foundation Models.

Methods	Publication	UCM-Captions
		CIDEr
MiniCPM-V	arXiv24	0.000
MiniGPT-v2	arXiv23	**16.282**
LLaVa-1.5	NIPS24	0.004
Qwen-VL-Chat	arXiv23	12.992
Florence-2-L	CVPR24	<u>13.844</u>
Sphinx	arXiv23	0.000
Geochat	CVPR24	0.288
LHRS-Bot	ECCV24	8.365
RemoteSAM	-	12.370

Table 12. Comparison with Foundation Models of Object Detection with horizontal bounding box.

Methods	Publication	DIOR	iSAID	DOTAv2
		$AP_{50}(\%)$		
MiniGPT-v2	arXiv23	9.430	2.54	1.62
Florence-2-L	CVPR24	26.98	16.67	12.25
Qwen-VL-chat	arXiv23	15.81	4.29	2.97
Sphinx	arXiv23	0.47	0.11	0.05
Falcon	arXiv25	<u>56.65</u>	<u>33.85</u>	**27.04**
RemoteSAM	-	**62.74**	**34.41**	<u>20.17</u>

Table 13. Object Detection performance with oriented bounding box of RemoteSAM.

Methods	Publication	DIOR	iSAID	DOTAv2
		$AP_{50}(\%)$		
Falcon	arXiv25	**55.29**	<u>28.83</u>	**23.29**
RemoteSAM	-	<u>55.22</u>	**34.58**	<u>18.21</u>

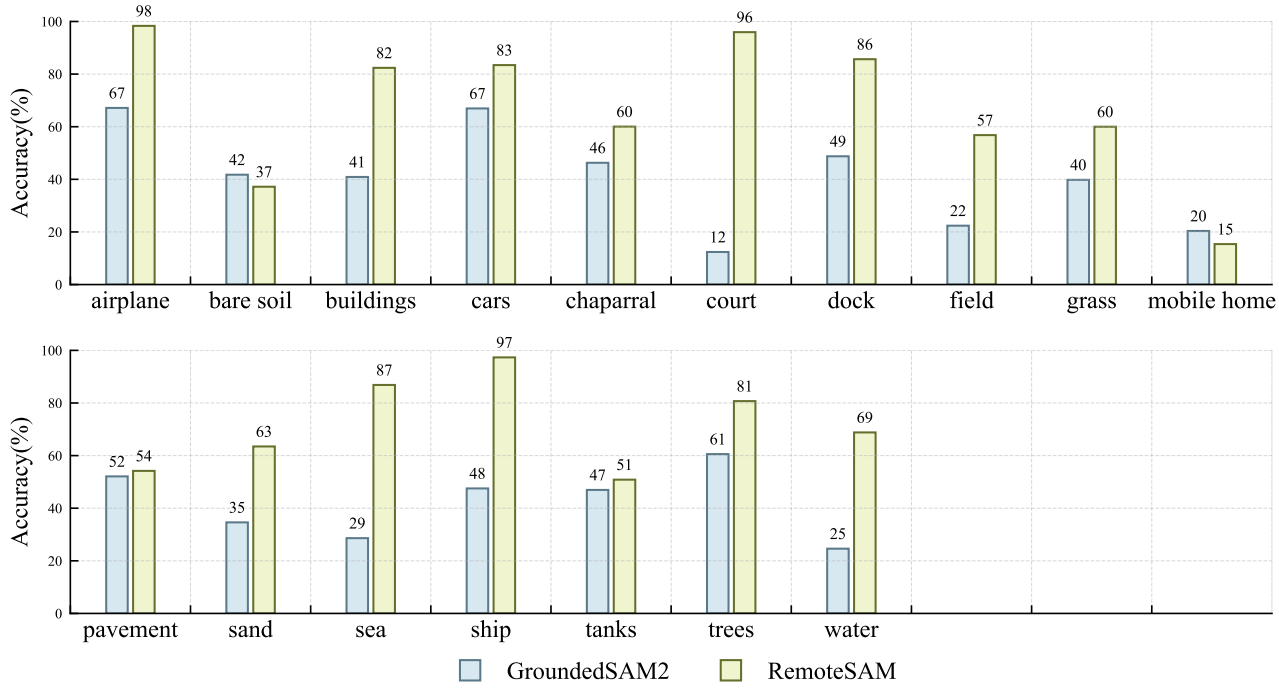


Figure 9. Detailed comparison with GroundedSAM2 on UCM_Multilabel in SATIN Complex Scenes Task

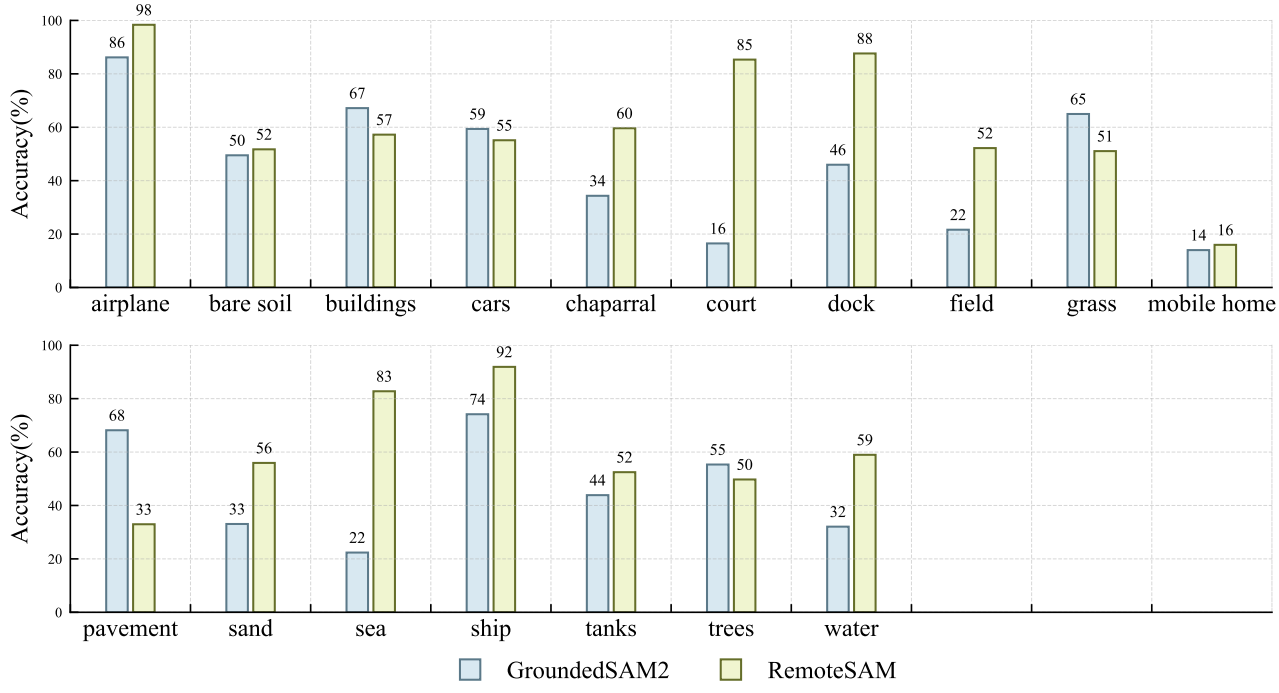


Figure 10. Detailed comparison with GroundedSAM2 on AID_Multilabel in SATIN Complex Scenes Task

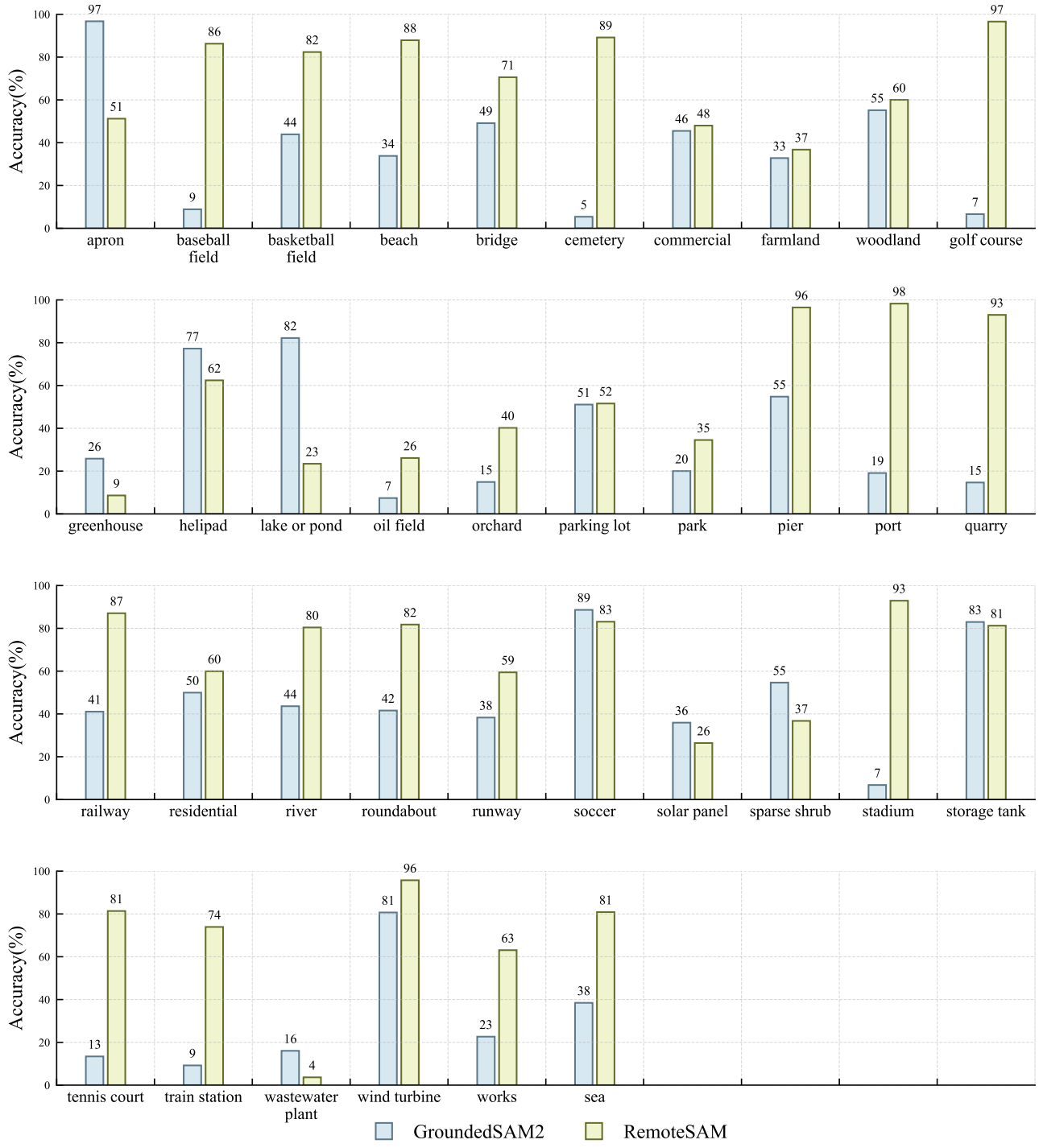


Figure 11. Detailed comparison with GroundedSAM2 on MultiScene in SATIN Complex Scenes Task

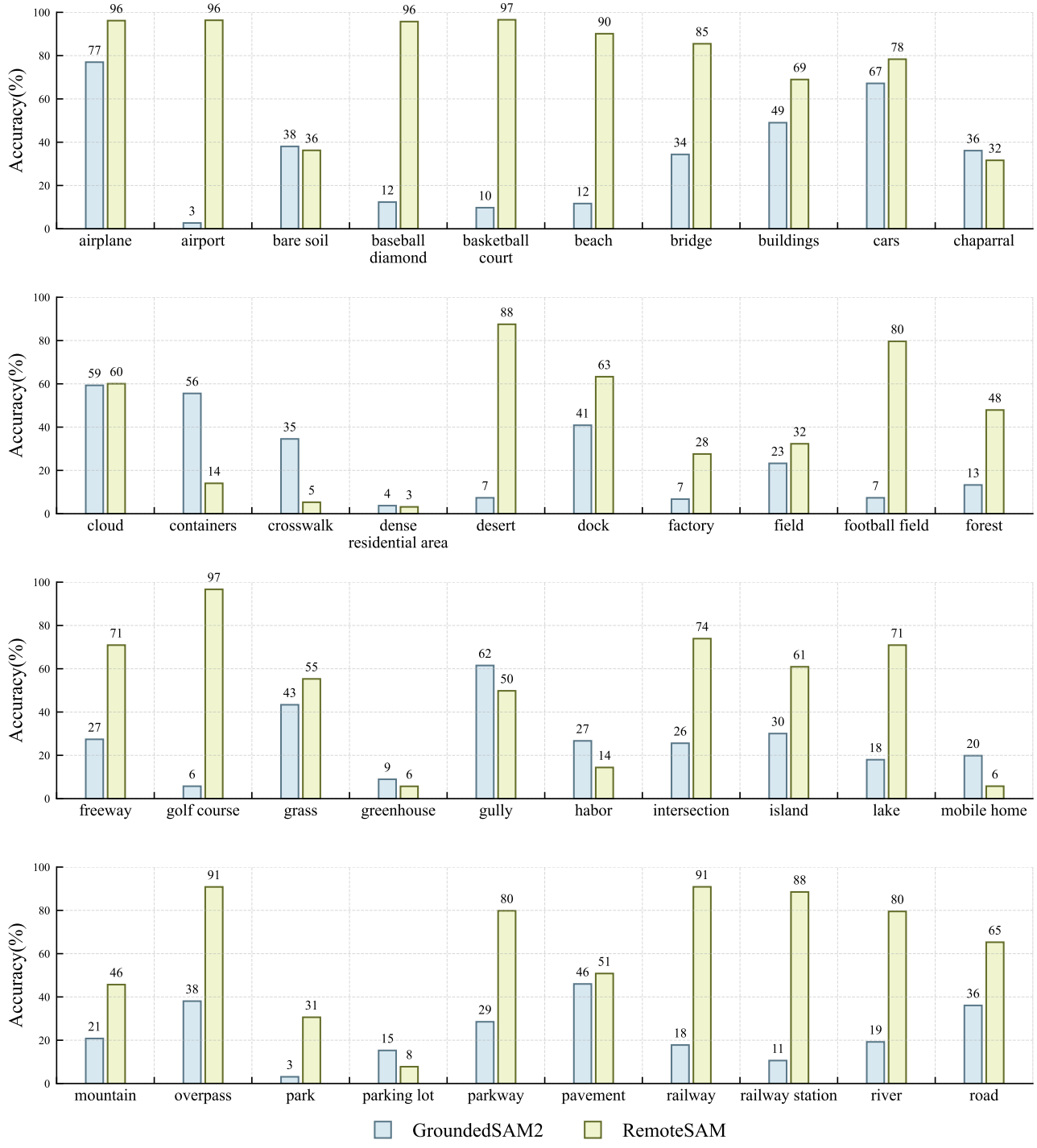


Figure 12. Detailed comparison with GroundedSAM2 on MLRSNet in SATIN Complex Scenes Task (Part I)

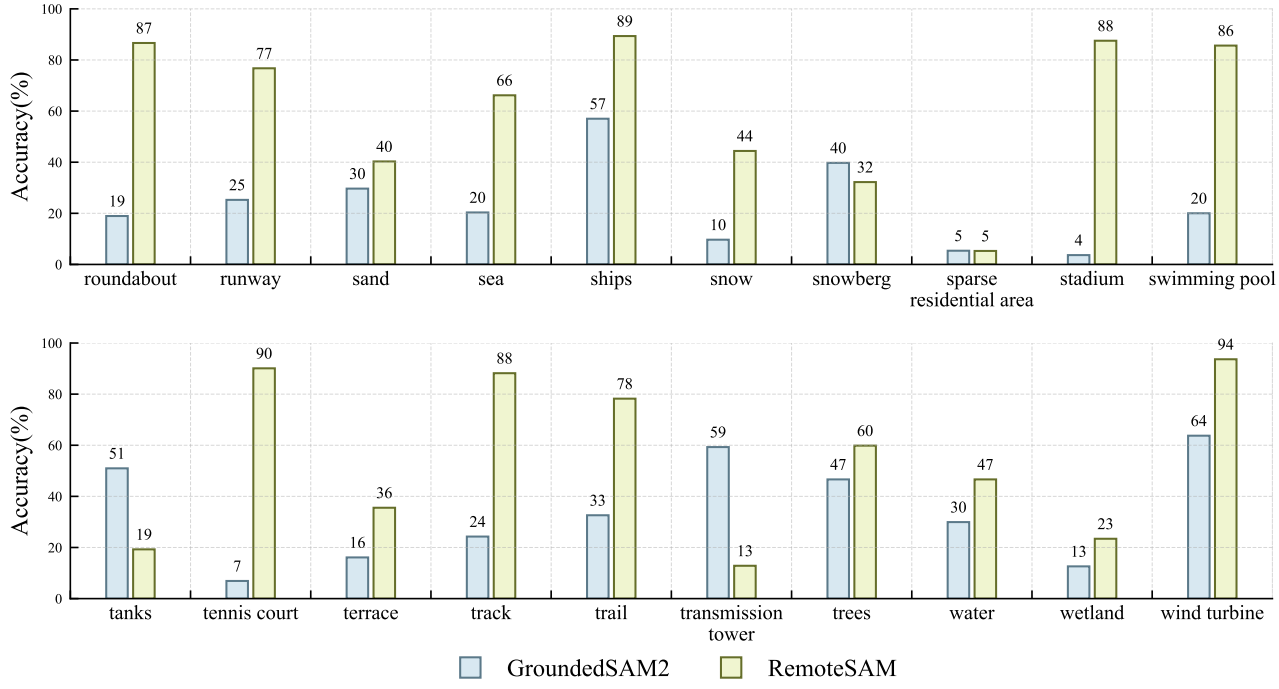


Figure 13. Detailed comparison with GroundedSAM2 on MLRSNet in SATIN Complex Scenes Task (Part II)

B. Qualitative results

In this section, we present the visualized results for each task as follows.

B.1. Task1: Referring Segmentation

Figure 14. Task1: Referring Segmentation

B.2. Task2: Semantic Segmentation

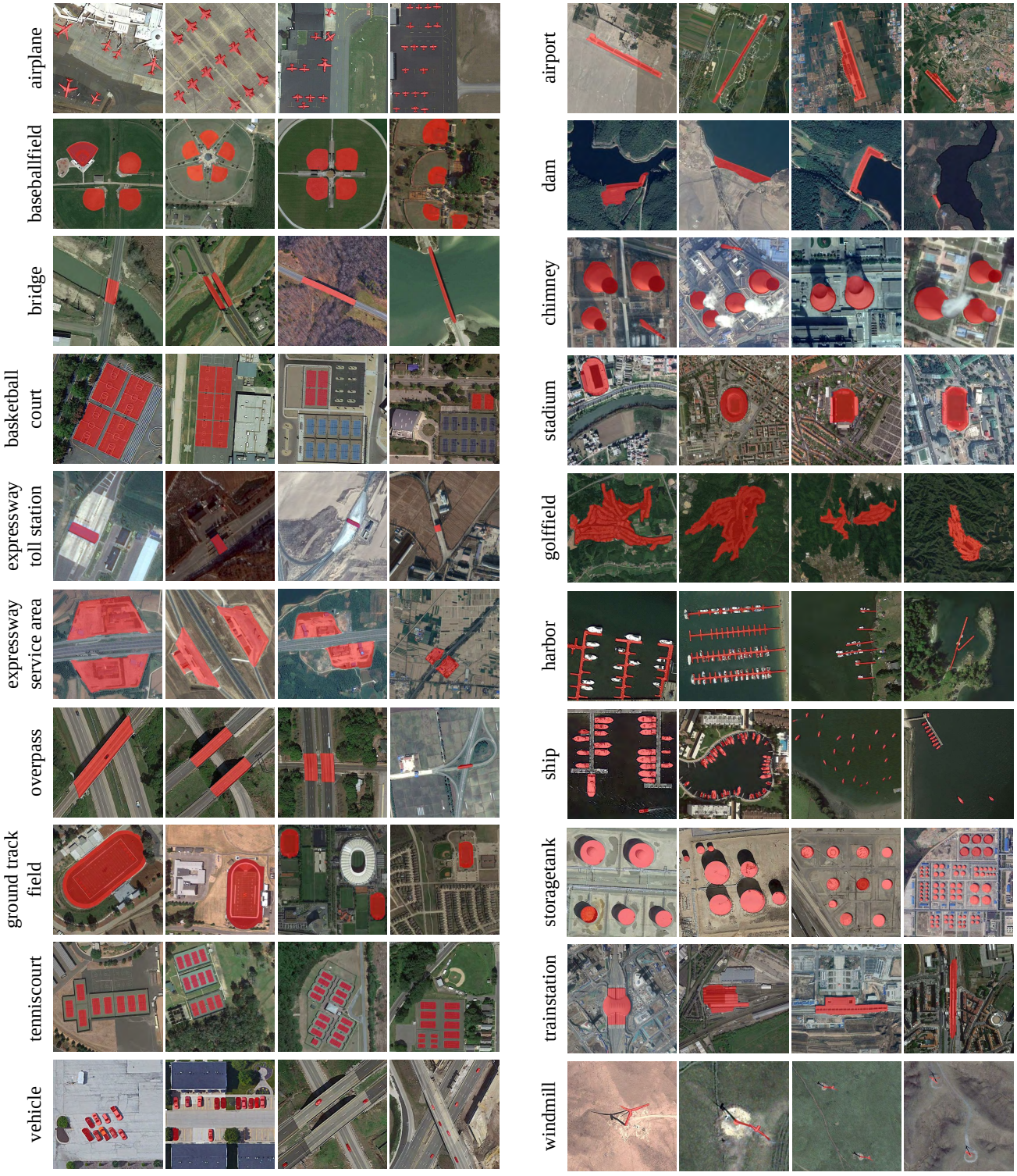


Figure 15. Task2: Semantic Segmentation

B.3. Task3: Object Detection

Figure 16. Task3: Object Detection

B.4. Task4: Object Counting

Figure 17. Task4: Object Counting

B.5. Task5: Visual Grounding

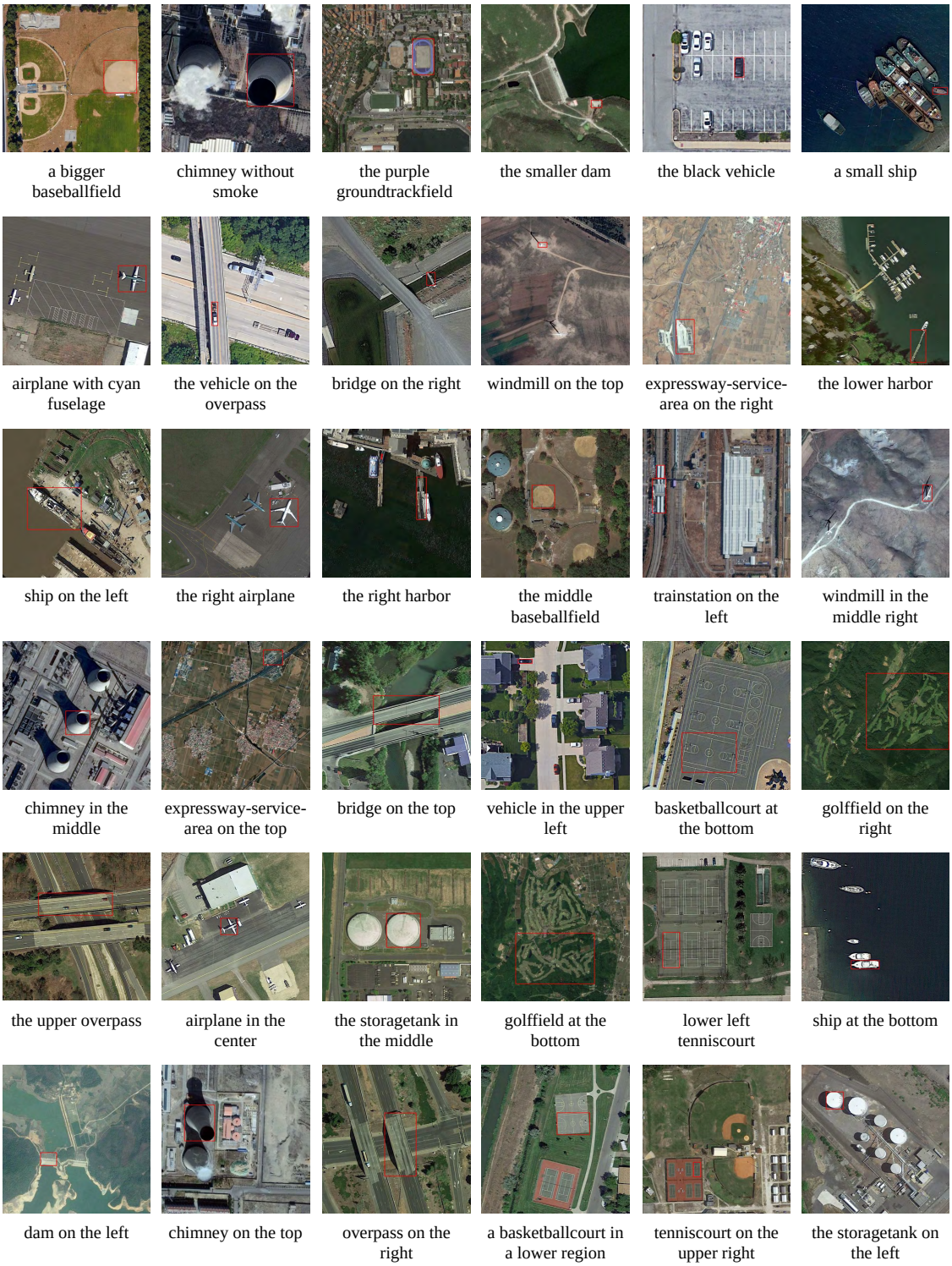
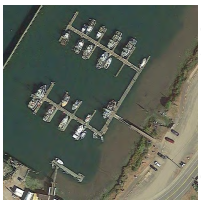


Figure 18. Task5: Visual Grounding

B.6. Task6: Multi-Label Classification

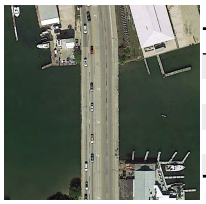
	Label	Score
√	baseballfield	0.517
√	tenniscourt	0.515
×	vehicle	8.5e-14
×	stadium	2.2e-14
×	airport	9.4e-15

	Label	Score
√	overpass	0.523
√	vehicle	0.502
×	bridge	1.8e-14
×	stadium	1.2e-14
×	storagetank	6.6e-15



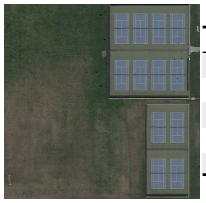
	Label	Score
√	harbor	0.525
√	ship	0.524
√	vehicle	0.501
×	overpass	2.1e-14
×	stadium	1.2e-14

	Label	Score
√	airplane	0.533
√	vehicle	0.503
×	ship	5.0e-14
×	trainstation	1.4e-14
×	windmill	2.9e-15



	Label	Score
√	bridge	0.572
√	harbor	0.509
√	vehicle	0.504
√	ship	0.503
×	storagetank	6.1e-14

	Label	Score
√	ship	0.578
√	harbor	0.555
×	bridge	9.4e-14
×	golffield	8.6e-15
×	overpass	5.8e-15



	Label	Score
√	tenniscourt	0.564
×	vehicle	9.3e-14
×	golffield	3.6e-14
×	dam	1.7e-14
×	windmill	7.8e-15

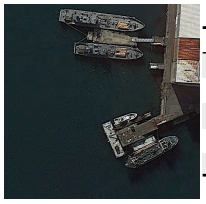
	Label	Score
√	chimney	0.569
√	storagetank	0.510
×	trainstation	5.8e-14
×	harbor	2.8e-14
×	dam	2.0e-14

	Label	Score
√	basketballcourt	0.549
√	tenniscourt	0.522
×	windmill	6.0e-14
×	bridge	2.2e-14
×	harbor	1.6e-14

	Label	Score
√	groundtrackfield	0.654
√	baseballfield	0.511
×	vehicle	5.9e-14
×	stadium	2.6e-14
×	ship	1.9e-14

	Label	Score
√	airplane	0.522
√	vehicle	0.501
×	stadium	8.7e-14
×	windmill	2.0e-14
×	ship	1.0e-14

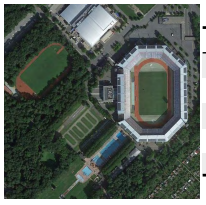
	Label	Score
√	windmill	0.501
×	golffield	2.8e-14
×	overpass	1.5e-14
×	airport	1.2e-14
×	bridge	3.8e-15



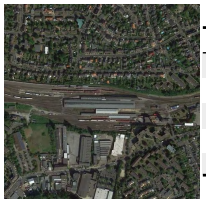
	Label	Score
√	ship	0.539
√	harbor	0.517
×	basketballcourt	5.3e-14
×	overpass	1.7e-14
×	dam	1.6e-14

	Label	Score
√	golffield	0.570
×	harbor	2.1e-14
×	bridge	2.0e-14
×	overpass	1.8e-14
×	airport	1.4e-14

	Label	Score
√	baseballfield	0.526
√	basketballcourt	0.516
√	vehicle	0.503
×	airport	7.7e-14
×	tenniscourt	6.6e-14



	Label	Score
√	stadium	0.558
√	groundtrackfield	0.533
√	vehicle	0.500
×	airport	1.2e-13
×	bridge	1.8e-14



	Label	Score
√	trainstation	0.518
√	vehicle	0.501
×	tenniscourt	3.1e-14
×	ship	2.2e-13
×	basketballcourt	1.5e-14

	Label	Score
√	basketballcourt	0.544
×	bridge	3.1e-13
×	golffield	2.7e-13
×	ship	5.3e-14
×	baseballfield	5.2e-14

	Label	Score
√	airport	0.511
×	vehicle	2.2e-14
×	storagetank	1.9e-14
×	bridge	7.8e-15
×	tenniscourt	7.3e-15

	Label	Score
√	chimney	0.549
√	storagetank	0.502
×	tenniscourt	6.4e-13
×	vehicle	1.3e-13
×	overpass	5.5e-14

	Label	Score
√	airplane	0.517
√	vehicle	0.505
×	trainstation	1.7e-14
×	storagetank	5.4e-15
×	bridge	5.0e-15

Figure 19. Task6: Multi-Label Classification

B.7. Task7: Image Classification

Figure 20. Task7: Image Classification

B.8. Task8: Image Caption

Figure 21. Task8: Image Caption

C. Ablation studies

Backbone Ablation: In Tab. 14, we investigate the effects of the text and image backbones through ablation experiments conducted on the RRSISD dataset with RemoteSAM. The combination of BERT as the text encoder and Swin-Base as the image encoder yields the highest performance, achieving an oIoU of 76.21% and an mIoU of 64.79%.

In contrast, substituting the text encoder with Transformer while retaining Swin-Base results in a slight decrease in performance, with scores of 75.51% for oIoU and 63.00% for mIoU. This indicates that while Transformer remains competitive, BERT provides a marginally better contextual representation for Referring Expression Segmentation. Moreover, when ConvNext-B was employed as the image encoder, the performance further declined across both text encoders, suggesting that ConvNext-B may not capture the spatial hierarchies as effectively as Swin-Base in the context.

As a result, we select BERT as the text encoder and Swin-Base as the image encoder for our model, furnishing a powerful foundational computation unit for task unification.

Table 14. Ablation study on Different Backbone

Text Encoder	Image Encoder	RRSISD	
		$oIoU(\%)$	$mIoU(\%)$
BERT	Swin-B	**76.21**	**64.79**
Transformer	Swin-B	<u>75.51</u>	<u>63.00</u>
BERT	ConvNext-B	73.77	61.75
Transformer	ConvNext-B	73.51	60.91

Effectiveness of CLIP Filtering: To validate the effectiveness of CLIP in filtering erroneous samples, we randomly sample 100 images from the filtered set and manually annotate their masks. As shown in the Tab. 15, the mIoU between the pseudo-labels and the manually annotated masks is 74.14%. This result indicates that our reserved pseudo-labels are accurate.

Table 15. Effectiveness of CLIP Filtering Strategy

Metrics	$Pr@0.5$	$Pr@0.6$	$Pr@0.7$	$Pr@0.8$	$Pr@0.9$	$oIoU(\%)$	$mIoU(\%)$
Pseudo-labels	85.00	79.00	71.00	55.00	34.00	66.37	74.14

Ablation of Multi-Label classification strategy : In Tab. 16, we explore different strategies for Multi-Label classification. For mask-level strategy, a label is judged as positive when the area of its corresponding mask exceeds an area threshold(default as 0). While for the prob-level strategy, labels are considered positive when their confidence scores obtained from pooling surpass a threshold(default as 0.5). As shown in the Table, the prob-level strategy significantly outperforms the mask-level approach on both datasets.

Table 16. Ablation of Multi-Label classification strategy

Strategy	DIOR	DOTAv2
	$Acc(\%)$	$Acc(\%)$
Mask-level	92.708	66.774
Prob-level	**94.042**	**75.752**

Balance factor analysis of classification: To ascertain the optimal balance factor values for classification, we conduct experiments by varying the balance factor λ within the lass-wise probability aggregation function (Eq.3). The experimental results are illustrated in Fig. 22. The results indicate that when λ is set to 0.5 and 1, Multi-Label Classification and Image Classification achieved the highest accuracy, respectively. Therefore, we adopt this set of parameters for our final test results.

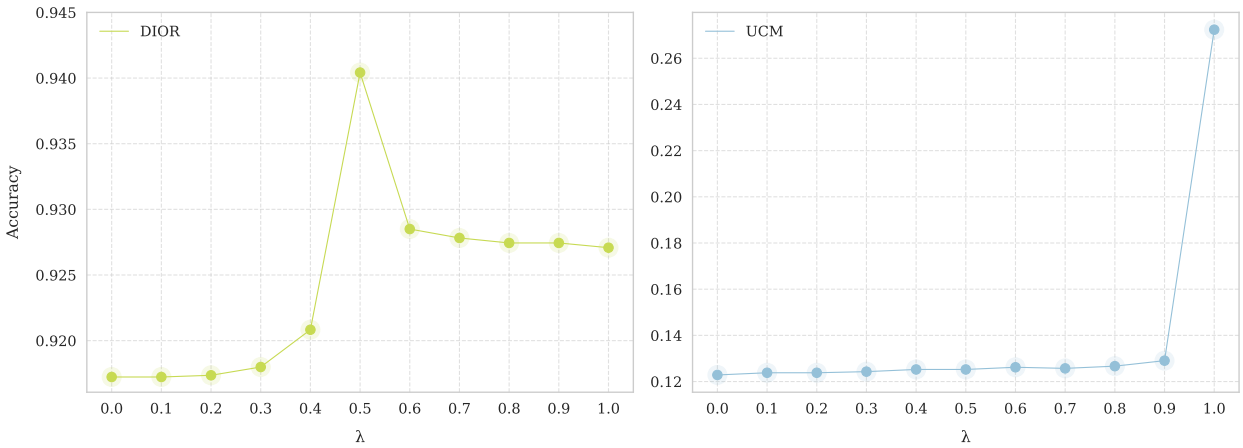


Figure 22. Balance factor analysis of classification

Qualitative examples of EPOC refinement in Object Detection: As illustrated in Fig. 23, the refinement by EPOC effectively resolves the limitations of the M2B strategy in processing adjacent targets.

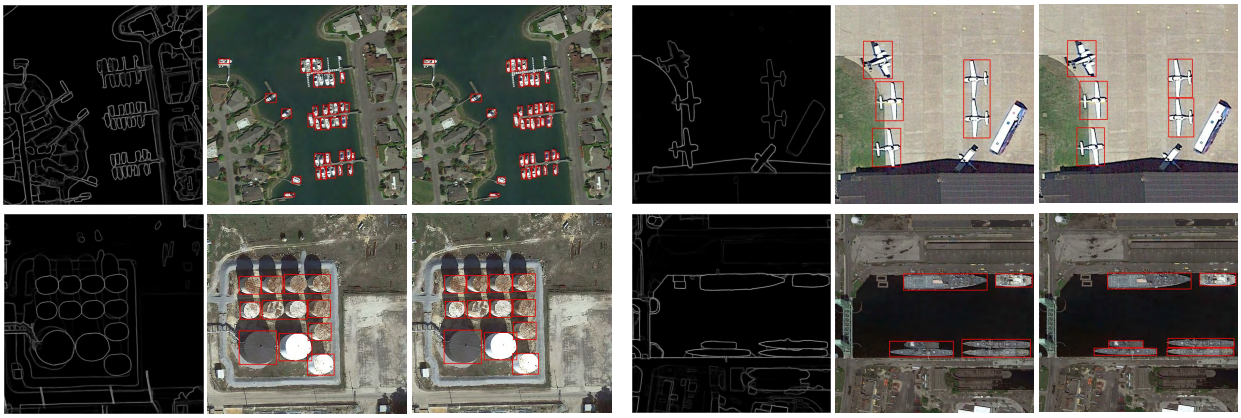


Figure 23. Qualitative examples of EPOC refinement in Object Detection

D. Examples of RemoteSAM-270k

Figure 24. Examples and categories of RemoteSAM-270k

E. Prompts Setting of Expressions Creation for Qwen2.5VL

Table 17. An Example implementation of extracting class-text pairs by prompting Qwen2.5VL.

Prompt for Extracting Class-Text Pairs

System Message:

Input:
- You will receive a detailed image caption of a remote sensing image.

Task Objective:
- Please extract all Object categories from the caption.
- Please generate a new concise description for each category.

Guidelines:
- Please focus on describing this single-class object and its attributes and spatial relation-ships.
- The caption must be brief, and no more than 20 words.
- The final output is in JSON format, with these categories as keys and corresponding descriptions as values.

User:

Detailed Caption: This image is an aerial view of an airport terminal and its surroundings. There are numerous gates around the terminal, each connected to the main building via jet bridges. The apron area is extensive, with many aircraft parked at the gates...

Assistant:

Json Format Output:

{
“aircraft”: “Multiple aircraft parked on the apron.”
“jet bridges”: “Multiple jet bridges connecting gates to the terminal.”
...
}