

# Co-Training Vision Language Models for Remote Sensing Multi-task Learning

Qingyun Li<sup>1\*</sup>, Shuran Ma<sup>3\*</sup>, Junwei Luo<sup>5\*</sup>, Yi Yu<sup>6\*</sup>, Yue Zhou<sup>4†</sup>, Fengxiang Wang<sup>7</sup>,  
 Xudong Lu<sup>8</sup>, Xiaoxing Wang<sup>2</sup>, Xin He<sup>1</sup>, Yushi Chen<sup>1✉</sup>, Member, IEEE,  
 Xue Yang<sup>2‡</sup>, Member, IEEE, Junchi Yan<sup>2</sup>, Senior Member, IEEE

<sup>1</sup>*Harbin Institute of Technology* <sup>2</sup>*Shanghai Jiao Tong University* <sup>3</sup>*Xidian University*

<sup>4</sup>*East China Normal University* <sup>5</sup>*Wuhan University* <sup>6</sup>*Southeast University*

<sup>7</sup>*National University of Defense Technology* <sup>8</sup>*Chinese University of Hong Kong*

\* Equal Contribution    ✉ Corresponding Author    † Project Leader

Code: <https://github.com/VisionXLab/RSCoVLM>

Data: <https://huggingface.co/datasets/Qingyun/remote-sensing-sft-data>

**Abstract**—With Transformers achieving outstanding performance on individual remote sensing (RS) tasks, we are now approaching the realization of a unified model that excels across multiple tasks through multi-task learning (MTL). Compared to single-task approaches, MTL methods offer improved generalization, enhanced scalability, and greater practical applicability. Recently, vision language models (VLMs) have achieved promising results in RS image understanding, grounding, and ultra-high-resolution (UHR) image reasoning, respectively. Moreover, the unified text-based interface demonstrates significant potential for MTL. Hence, in this work, we present RSCoVLM, a simple yet flexible VLM baseline for RS MTL. Firstly, we create the data curation engine, including data acquisition, offline processing and integrating, as well as online loading and weighting. This data engine effectively addresses complex RS data environment and generates flexible vision-language conversations. Furthermore, we propose a unified dynamic-resolution strategy to address the diverse image scales inherent in RS imagery. For UHR images, we introduce the Zoom-in Chain mechanism together with its corresponding dataset, LRS-VQA-Zoom. The strategies are flexible

This work was supported by National Natural Science Foundation of China under the Grant 62371169 and 62506229, and Natural Science Foundation of Shanghai under 25ZR1402268.

Q. Li, X. He, and Y. Chen are with the School of Electronics and Information Engineering, Harbin Institute of Technology, Harbin 150001, China (e-mails: 21B905003@stu.hit.edu.cn; hexin1@hit.edu.cn; chenyushi@hit.edu.cn). (Corresponding author: Yushi Chen.)

S. Ma is with the School of Telecommunications Engineering, Xidian University, Xi'an 710071, China (e-mails: shrma@stu.xidian.edu.cn).

J. Luo is with the School of Remote Sensing and Information Engineering, Wuhan University, Wuhan 430072, China (e-mail: luojunwei@whu.edu.cn).

Y. Yu is with the School of Automation, Southeast University, Nanjing 210096, China (e-mail: yuyi@seu.edu.cn).

F. Wang is with the College of Computer Science and Technology, National University of Defense Technology, Changsha 410005, China (e-mail: wfx23@nudt.edu.cn).

X. Lu is with the Department of Electronic Engineering, The Chinese University of Hong Kong, Hong Kong SAR 999077, China (e-mail: luxudong@link.cuhk.edu.hk).

X. Wang is with the School of Computer Science, Shanghai Jiao Tong University, Shanghai 200240, China (e-mail: figure1\_wxx@sjtu.edu.cn).

Y. Zhou is with the School of Geospatial Artificial Intelligence, East China Normal University, Shanghai 200241, China (e-mail: yzhou@geoai.ecnu.edu.cn).

X. Yang is with the School of Automation and Intelligent Sensing, Shanghai Jiao Tong University, Shanghai 200240, China (e-mail: yangxue-2019-sjtu@sjtu.edu.cn).

J. Yan is with the School of Artificial Intelligence, Shanghai Jiao Tong University, Shanghai 200030, China (e-mail: yanjunchi@sjtu.edu.cn).

and effectively mitigate the computational burdens. Additionally, we significantly enhance the model's object detection capability and propose a novel evaluation protocol that ensures fair comparison between VLMs and conventional detection models. Extensive experiments demonstrate that RSCoVLM achieves state-of-the-art performance across diverse tasks, outperforming existing RS VLMs and even rivaling specialized expert models. All the training and evaluating tools, model weights, and datasets have been fully open-sourced to support reproducibility. We expect that this baseline will promote further progress toward general-purpose RS models.

**Index Terms**—vision-language model, remote sensing, multi-task learning

## I. INTRODUCTION

EARTH observation systems have acquired extensive remote sensing (RS) data, necessitating the development of automated RS image interpretation techniques [1]. The emergence of artificial general intelligence has inspired researchers in the RS community to develop versatile agents capable of performing multiple tasks, such as scene classification, visual question answering, and object detection [2].

Most RS image processing methods typically train a specifically-designed model on isolated datasets to achieve optimal performance on individual tasks. Due to the heterogeneity of data and model architecture, developing a unified model capable of handling multiple RS tasks, i.e., multi-task learning (MTL), remains challenging [3].

MTL provides several advantages for RS applications. First, a single MTL model with shared parameters can handle multiple tasks at once, unlike traditional task-specific models, which is closer to human perception. Second, by sharing knowledge across tasks, MTL mitigates the shortage of annotated data and reduces overfitting on individual tasks. Third, MTL learns joint representations that capture correlations among tasks, improving generalization. RS foundation models also benefit by obtaining consistent representations through pre-training on upstream tasks and fine-tuning on various downstream tasks. Overall, MTL helps advance RS foundation models by expanding pre-training tasks and enhancing cross-task learning [4].

	<u>Resolution</u>	⇒	<u>Architecture</u>	⇒	<u>Supported Tasks</u>
<b>Vision MTL</b> e.g., <i>RSCoTr</i> , <i>MTP</i> , <i>SM3Det</i> , ...	?	?	?	?	<i>Det.</i> <i>Grd.</i> <i>Desc.</i> <i>Cls.</i>
	<b>Regular Images</b> Up to 1024×1024		<b>Task Specific</b> Decoders for Each Task		😊 😐 😐 😊
<b>Detection VLMs</b> e.g., <i>LMMRotate</i> , ...	?		😊		😊 😐 😐 😐
<b>Regular VLMs</b> e.g., <i>VHM</i> , <i>GeoChat</i> , ...	?	<b>Regular Images</b> Up to 1024×1024	<b>VLM Based</b> Unified Textual Outputs		😐 😊 😊 😊
<b>Grounded VLMs</b> e.g., <i>GeoGround</i> , ...					😐 😊 😐 😐
<b>UHR VLMs</b> e.g., <i>GeoLLaTA8k</i> , ...	?	<b>UHR Images</b> Ultra-High-Resolution	<b>VLM Based</b> Unified Textual Outputs		😐 😐 😊 😊
<b>Unified RS Multi-task VLM (Ours)</b>	😊	<b>Regular / UHR</b> Any Input Resolution	<b>VLM Based</b> Unified Textual Outputs		😊 😊 😊 😊
					<b>Versatile</b> Multi-task VLM

Fig. 1. Comparisons with existing MTL methods across the resolutions of input images, network architectures, and supported tasks (i.e., detection, grounding, description, and classification).

Transformer [5] has demonstrated remarkable flexibility and generalization capabilities across various domains, including computer vision [6], natural language processing, speech processing, and remote sensing data analysis [7]. This progress has brought the goal of a unified multimodal and multi-task architecture increasingly within reach [8]. Consequently, vision language models (VLMs), that bridge the gap between the two modalities by learning from vast amounts of paired [9] and interleaved image-text data [10], have been proposed and become the most commonly adopted foundation model paradigm in the multimodal domain [11].

In this study, we focus on generative VLMs, also named multimodal language models. These models are typically constructed upon vision and language foundation models, enabling them to process visual inputs and effectively interpret textual instructions. By harnessing the capabilities of powerful pre-trained foundation models and leveraging a versatile text interface, VLMs are positioned as a crucial element in the progression toward the unified MTL [12].

We consider that VLMs represent an ideal paradigm for RS MTL. Firstly, the textual interface of VLMs provides a unified representation for diverse task objectives, because the outputs of different RS tasks, such as classification, grounding, captioning, or question answering, can all be expressed in text form. Secondly, instruction tuning has demonstrated that VLMs can generalize beyond the tasks seen during training [13], enabling them to handle novel or composite tasks through in-context learning [11]. Finally, with sufficiently strong foundational capabilities, VLMs offer the potential to evolve toward more autonomous RS agents, where task reasoning and workflow design can be accomplished within a single, coherent framework.

In the RS community, MTL has been preliminarily explored, including several attempts leveraging VLMs. Nevertheless, existing approaches still exhibit notable limitations. Fig. 1 summarizes the key differences among representative paradigms.

Early RS MTL approaches were typically designed for multiple pure-vision tasks [4], [14], [15], such as classification, segmentation, and detection. These methods generally adopt a shared feature extraction backbone with task-specific

output heads. With carefully crafted training strategies, their performance on individual benchmarks is comparable to that of expert models trained on the specific dataset. However, they suffer from limited scalability and architectural rigidity. As the number of tasks increases, the heterogeneous design of multiple heads makes optimization increasingly difficult and less robust. Consequently, this paradigm struggles to scale up, resulting in insufficient model generalization. Nevertheless, when deployed on resource-constrained platforms such as satellites, this kind of MTL model remains highly valuable for its computational and storage efficiency.

As general-purpose VLMs increasingly exhibit early signs of the universal model, they have emerged as a scalable paradigm for MTL. In the RS domain, several studies have explored VLM-based MTL. However, their investigations into unified and generalizable paradigms remain limited: The regular VLMs focus primarily on language-centric description tasks such as image captioning and visual question answering, where text descriptions are synthesized for RS images to enable semantic understanding [16]–[18]. Others extend VLMs to purely visual tasks such as visual grounding and object detection, leveraging the flexible language interface of VLMs to learn from abundant localization annotations and achieve precise detection capabilities [19], [20]. In addition, several approaches target ultra-high-resolution (UHR) RS image reasoning, often employing token pruning to alleviate the computational burden caused by extremely large inputs [21], [22].

These studies collectively highlight the great potential of VLMs for RS MTL, yet each remains constrained within a limited scope. As shown in Fig. 1, the first four types of works focus mainly on tasks involving regular images (images with regular resolutions) [16]–[20], whereas the fifth is specialized for UHR scenarios [21], [22]. The detection VLMs [20] and grounded VLMs [19] excel at spatial grounding but pays little attention to semantic understanding, while the regular VLMs [16]–[18] rarely explore the crucial object detection capabilities which is essential for RS image analysis. Hence, a unified framework that addresses these limitations in an integrated MTL setting is still lacking.

In this paper, we present a novel foundation model named **RSCoVLM** (**R**emote **S**ensing **C**ooperatively-trained **V**sion **L**anguage **M**odel). We cooperatively train (co-train) it for multiple tasks in a unified framework, that handles the following problems.

Firstly, the large-scale multi-task data must be curated to enable effective MTL. However, RS data are inherently complex, often exhibiting inconsistencies in format, noisy annotations, and heterogeneous bounding box definitions. Therefore, careful data curation is required to construct a well-organized and sustainable data environment for model training.

Secondly, we need to address the challenge of diverse input sizes of RS images. The classification task often uses size with a few megapixels, such as  $256 \times 256$ . Common object detection models typically use input sizes such as  $512 \times 512$ ,  $800 \times 800$ , or  $1024 \times 1024$ . However, UHR images can have widths and heights exceeding 4,000 pixels. Therefore, a dynamic resolution strategy is required, along with efficient

and highly compatible solutions for UHR scenarios.

Finally, previous VLMs have shown limited capability in object detection tasks. They either perform only sparse visual grounding [17], [19], provide detection results for a single category [23], or are evaluated leniently under low IoU thresholds [24]. However, aerial detection is particularly challenging due to issues such as dense object distribution, which places high demands on the visual input resolution and output sequence length of MLLMs. Moreover, VLMs cannot directly output the confidence scores of predicted objects, making it difficult to fairly compare them with traditional models using commonly-used evaluation metrics.

To make RSCoVLM a competitive MTL baseline, we address the aforementioned challenges, respectively. We firstly create a data curation engine, compromising the acquisition of raw data, the offline processing and integration, as well as online loading and weighting. Moreover, we propose a dynamic resolution strategy that enables the model to simultaneously learn from images of various sizes. To further enhance the reasoning performance on UHR images, we propose the Zoom-in Chain strategy, which mimics how humans reason over UHR images. We also construct a corresponding dataset, LRS-VQA-Zoom, to specifically strengthen this capability. Additionally, we apply VLMs to object detection and propose a fair evaluation method that does not rely on confidence thresholds. Based on this, our RSCoVLM is validated as the first VLM that achieves performance comparable to traditional models on dense aerial detection task.

We evaluate RSCoVLM on multiple tasks across various benchmarks, achieving state-of-the-art performance in all of them. Our unified MTL framework greatly improves the model's generalization ability, scalability, and usability.

To ensure transparency and reproducibility, we fully open-source all details of this work, including the codes, model weights, data folder. We will continuously maintain the open-source resources and update them with our latest research progress, aiming to build a user-friendly platform for the community.

The main contributions are summarized as follows:

- 1) We present RSCoVLM, a fully open-sourced VLM baseline for RS MTL. The experiment show that our model achieves leading performance across benchmarks of various datasets and tasks.
- 2) We develop the universal framework for RS MTL based on VLM and create the data curation engine to facilitate unified training across multi datasets of various RS tasks.
- 3) We proposed a dynamic resolution strategy for RS, along with the Zoom-in Chain strategy and the LRS-VQA-Zoom dataset to further enhance the model's reasoning ability on UHR images.
- 4) We develop the auto-regressive aerial detection method for RS VLMs and propose an evaluation metric that enables a fair comparison between RS VLM and conventional methods.

This manuscript is an extended and improved version of our conference paper [20] published in IGARSS 2025, which only investigate VLMs for detection tasks. The autoregressive object detection scheme in Section III-E is primarily derived

from the conference version. Building upon it, we not only refine detection details in Section III-E1 but also further upgrade the VLMs with unified multi-task learning, accompanied by additional methods, models, and experimental results.

## II. RELATED WORKS

### A. General-purpose Vision Language Models

Early works such as VisualGPT [25], BLIP-2 [26], and Flamingo [11] explored different ways of integrating visual features with large language models or training joint image-text encoders, showing improved multimodal reasoning and understanding.

Later instruction-tuned frameworks, including LLaVA [13], MiniGPT-4 [27], and InstructBLIP [28], further enhanced interactive comprehension by fine-tuning LLMs with visual-text instructions. Lightweight adaptation methods such as LLaMA-Adapter V2 [29] and SPHINX [30] improved efficiency through visual adapters and zero-shot attention fusion, reducing the cost of multimodal alignment.

In parallel, the VisionLLM series [31], [32] unified vision-centric tasks under the LLM paradigm, enabling open-ended reasoning over diverse visual inputs. More recent large-scale MLLMs, including PaLI-X [33], MiMo-VL [34], InternVL [12], CogVLM [35], and the Qwen-VL series [36], [37], further integrate vision and language through end-to-end pretraining on massive multimodal data and scalable architectures. These models show improved visual grounding, OCR, and cross-domain reasoning, representing a shift from adapter-based fusion toward deeply coupled vision-language modeling. Collectively, these advances lay the foundation for adapting MLLMs to specialized domains such as remote sensing, where complex visual semantics and open-ended reasoning are required.

### B. Remote Sensing Vision Language Models

Recently, integrating vision-language models into remote sensing (RS) has attracted growing attention, giving rise to several domain-specific VLMs. GeoChat [17] pioneered this direction as the first RS-oriented VLM, addressing multiple optical imagery tasks via conversational interaction. EarthGPT [23] introduced a unified multimodal framework for multi-source RS data and diverse vision-language tasks. LHRS-Bot [18] leverages multi-level vision-language alignment and curriculum learning for RS image understanding.

Beyond static image interpretation, recent works explore temporal and fine-grained understanding. TEOChat [38] supports temporal Earth observation imagery and instruction following over sequential frames. SkySenseGPT [24] extends instruction tuning to fine-grained RS comprehension, achieving strong performance on public datasets and complex comprehension tasks. VHM [16] demonstrates capabilities on tasks such as building vectorization, multi-label classification, and honest question answering.

### C. Remote Sensing Multi-task Learning

While recent advancements in remote sensing VLMs have enabled versatile multi-task capabilities, their general-purpose

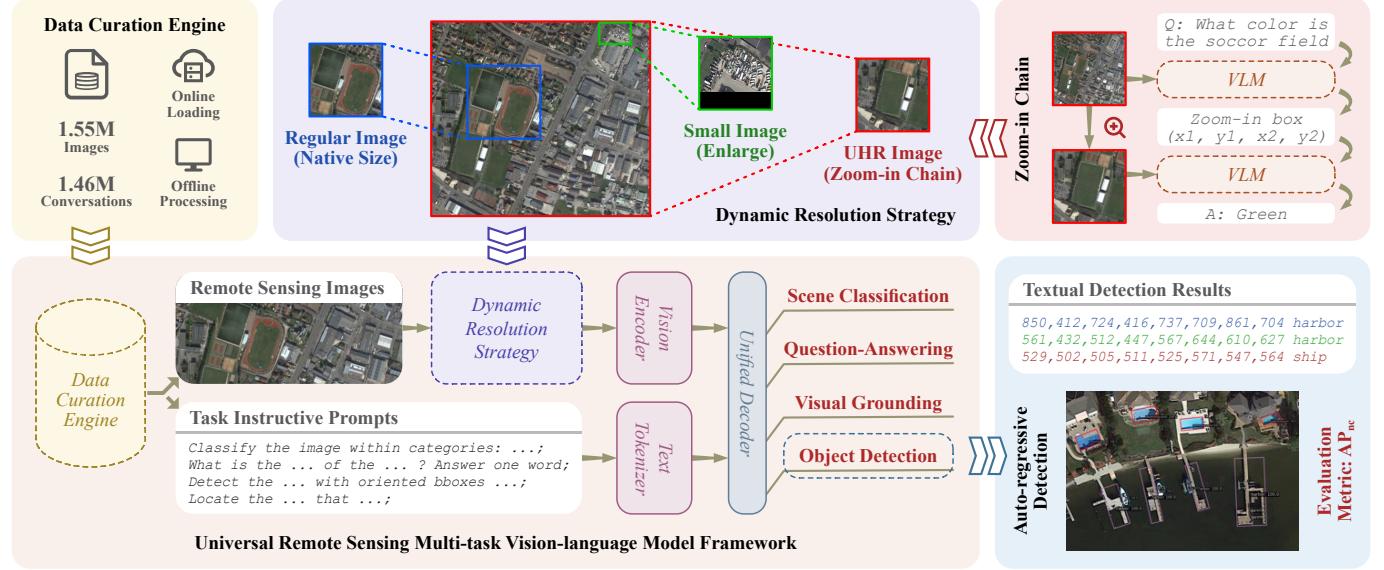


Fig. 2. Overall schematic diagram of the proposed method. The overall RS MTL framework based on VLM is presented in Section III-A. The data curation engine is introduced in Section III-B. The dynamic resolution strategy is proposed in Section III-C. We introduce the proposed Zoom-in Chain strategy and the corresponding LRS-VQA-Zoom dataset in Section III-D. Finally, we describe the aerial detection scheme and propose the fair metric AP<sub>nc</sub> in Section III-E.

architectures often implicitly handle task interdependencies, potentially overlooking the intrinsic challenges of multi-task optimization, such as task interference and imbalance.

One line of research focuses on leveraging shared representations for synergistic task pairing. For example, several studies [39]–[42] combine semantic segmentation with change detection from bi-temporal imagery, showing that joint learning enhances feature sharing and reduces redundancy. Another direction jointly models geometric and semantic information, such as height estimation with semantic segmentation [43], [44], demonstrating gains over single-task baselines.

Beyond specific task pairs, generalized MTL frameworks have been proposed. RSCoTr [4] performs classification, segmentation, and detection simultaneously, illustrating the potential of unified RS analysis. Large-scale datasets like SatlasPretrain [45] with multiple annotation modalities facilitate advanced MTL model development. SM3Det [15] uses a mixture-of-experts structure for multi-modal detection of horizontal and rotated bounding boxes. EarthDial [46] leverages multiple multi-task decoders to transfer knowledge across diverse tasks, enriching shared feature learning.

### III. METHOD

#### A. The Universal RS Multi-task Framework

As shown in Fig. 2, we propose a universal framework for RS MTL based on VLM. The model follows a popular VLM paradigm. It uses a vision encoder and text tokenizer to process images and text inputs, respectively. The unified decoder based on a language model then process the bi-modal features and perform various tasks, such as RS image scene classification, question answering, captioning, grounding, and object detection.

Specifically, we develop a data curation engine consisting of data acquisition, offline processing, and online loading, which provides diverse images with textual prompts and golden

responses for model training. To enable the model supporting images of arbitrary sizes, we design the dynamic resolution strategy, which handles input images of small, regular and UHR sizes respectively. The proposed Zoom-in Chain is designed to further enhance reasoning on UHR RS images. The final model can perform multiple tasks simultaneously. With the proposed auto-regressive aerial detection method, the model can perform the challenging aerial detection.

For the language branch, the input text is tokenized into a sequence of indices, where each index  $i$  corresponds to a learnable embedding  $\mathbf{t}_i \in \mathbb{R}^D$ . The output sequence is then de-tokenized to produce the final textual response.

For the vision branch, a RS image is preprocessed (e.g., resized or dynamically rescaled) and encoded by a vision Transformer to obtain features  $\mathbf{F} \in \mathbb{R}^{N_I \times D_I}$ , where  $N_I$  and  $D_I$  denote the feature number and dimension. The prompt text is tokenized into  $N_t$  embeddings  $\mathbf{T}_t \in \mathbb{R}^{N_t \times D}$ . A bi-modal projection aligns visual embeddings with the language token space, generating  $N_v$  visual tokens  $\mathbf{T}_v \in \mathbb{R}^{N_v \times D}$ , with  $N_v \propto N_I$ . The language model input is:

$$\mathbf{T} = \text{concat}(\mathbf{T}_v, \mathbf{T}_t) \in \mathbb{R}^{(N_v + N_t) \times D}, \quad (1)$$

where  $\text{concat}(\cdot, \cdot)$  denotes token-wise concatenation.

During training, parameters  $\theta$  are optimized via next-token prediction using cross-entropy loss:

$$\mathcal{L} = - \sum_{j=1}^{|r|} P_j(\mathbf{r}, \mathbf{T}), \quad P_j(\mathbf{r}, \mathbf{T}) = \log P_\theta(\mathbf{r}_j | \mathbf{r}_{<j}, \mathbf{T}), \quad (2)$$

where  $\mathbf{r} = (\mathbf{r}_1, \dots, \mathbf{r}_T)$  is the response token sequence.

During inference, the model generates tokens auto-regressively until an end-of-sequence token is reached:

$$\mathbf{r}_j = \arg \max P_j(\mathbf{r}, \mathbf{T}) \quad \text{or} \quad \mathbf{r}_j \sim P_j(\mathbf{r}, \mathbf{T}), \quad (3)$$

where the first denotes deterministic decoding (e.g., greedy or beam search) and the second stochastic sampling (e.g., top- $k$ , nucleus sampling).

### B. Data Curation Engine

In contrast to conventional approaches that mainly conduct standardized evaluations on a single benchmark, this section highlights the crucial role of data curation in developing RS MTL models. Given the diversity and complexity of RS data—characterized by heterogeneous formats, noisy annotations, and inconsistent bounding box definitions. Hence, a well-curated data recipe is indispensable. To serve as a robust foundation for training RS multi-task VLMs, we design a data curation Engine, which is not a fixed dataset but a comprehensive and sustainable data framework encompassing the following three main parts.

1) *Data Acquisition*: In this work, the dataset was curated through three sequential stages. Initially, we collected data by following the data recipes of several representative open-source vision-language models. Specifically, we adopted the description-related subsets from the instruction tuning data of VHM [16] and GeoChat [17], which cover tasks such as image classification, captioning, and visual question answering. These tasks can be further decomposed into subtasks, including modality recognition and resolution estimation. The refGeo [19] dataset was employed as the main grounding data source, while temporal multi-image data were drawn from TEOChatlas [38]. To prevent degradation of the model’s general reasoning ability during continued training, we also incorporated a subset of general-purpose data sampled from LLaVA-OneVision’s recipe [47], including chart interpretation, optical character recognition, and so on. By following these open-source data recipes, we indirectly surveyed and integrated diverse data sources.

Subsequently, we analyzed the limitations of the collected data and expanded the dataset using task-specific training set. We observed that existing RS VLMs rarely address object detection, which is crucial for fine-grained perception in RS. To fill this gap, we incorporated the DOTA-v1.0 dataset [48], thereby enriching the model’s detection-related learning capabilities.

Finally, for abilities that could not be obtained from open datasets, we constructed a synthetic data pipeline to generate new annotations. To enable the model’s zoom-in chain capability, we curated large-scale RS images and synthesized image–region–question triples. The detailed construction process is described in Section III-D.

2) *Data Processing and Integrating*: Due to the diverse formats and task requirements of the collected datasets, as well as potential systematic noise, we performed additional offline preprocessing to integrate all data into our training framework.

We first removed all task descriptors, such as the “[grounding]”, “[refer]”, and “[identify]” tags used in previous works [16], [17]. These descriptors tag the specific tasks. However, in open-world scenarios or novel tasks, instructions are typically expressed in natural language rather than through fixed descriptor tokens. Therefore, we replaced these descrip-

tors with natural language prompts to better align with the real-world usage.

Next, we examined all bounding boxes in the datasets and categorized them into horizontal boxes, oriented boxes, and quadrilateral boxes. Their representations were then unified through consistent normalization and ordering to avoid any information mismatch. Corresponding prompts were designed for each box type. By default, horizontal boxes were used in grounding tasks, while quadrilateral boxes were adopted for detection tasks.

A unified data format was further established to standardize the integration. Conversational data followed the messages structure defined by OpenAI, object detection data were formatted according to the COCO convention, and grounding data adhered to the refGeo [19] schema.

Finally, we performed rule-based cleaning on systematic irregularities, such as removing redundant punctuation and spaces, and correcting typographical errors. For the Zoom-in Chain dataset, we applied tool-call formatting. The evaluation set was also processed in a similar manner to ensure consistency with the training data.

3) *Data Loading and Weighting*: After integration, the curated dataset was organized into multiple subset units. During training, we applied online preprocessing and dynamically controlled the sampling ratio of each subset. Consequently, the model was trained in a flexible and adaptive multi-task environment rather than on a fixed, pre-defined dataset.

We argue that the model should not rely solely on predefined prompts from the training stage. To enhance robustness, multiple agent prompts were designed for certain tasks, and one was randomly selected during training. For grounding and detection data, a unified formatting scheme was adopted. We also incorporated the JSON-based output format used in Qwen2.5-VL [37], accompanied by specific prompts, and randomly switched between standard and JSON outputs during training. In addition, a synonym replacement module was implemented to randomly substitute words with their synonyms, improving the model’s linguistic generalization. Standard data augmentation techniques, such as random resizing, were also applied to enhance multi-scale learning.

Each subset unit was assigned a sampling weight to guide data selection during training, analogous to controlling the flow rate of different ingredients in an automatic beverage dispenser. The sampling ratio is critical for multi-task learning: increasing the weight of more challenging tasks facilitates deeper learning, while adjusting the others helps mitigate catastrophic forgetting. In exploring optimal weighting strategies, we first conducted experiments with uniform ratios. Then, we increased the sampling proportion of tasks that underperformed relative to expectations. Finally, once all tasks reached or exceeded satisfactory performance, we fine-tuned the weights to achieve the best overall multi-task balance.

### C. Dynamic Resolution Strategy

Most existing RS VLMs (such as GeoChat [17], VHM [16], and GeoGround [19]) have the only fixed square input shape (such as  $336 \times 336$  or  $504 \times 504$ ). For each input image,

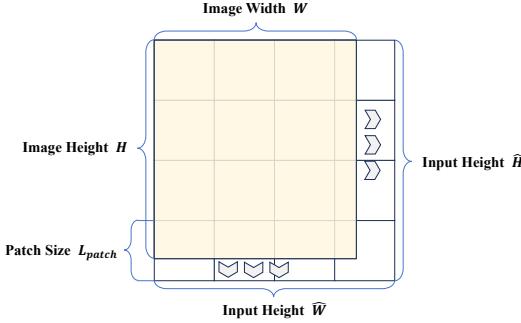


Fig. 3. Schematic diagram of the native resolution input.

they first pad the image to a square with zeros on the right or bottom, and then resize it to the input shape. Additionally, LRS-VQA [21] and GeoLLaVA-8k [22] scale the input size to  $2k \times 2k$  and  $8k \times 8k$ , respectively. They first cut a UHR RS image into slices of the fixed size and encode them into visual tokens. Then they prune the tokens to an amount comparable to the normal cases. In total, they pre-process the images only on a fixed shape or a small set of image shapes.

The proposed dynamic resolution strategy involves three interconnected aspects: supporting full-size input processing, scaling coordinate precision with input resolution, and curating training data to enhance learning across multiple resolutions.

1) *Full-scale Visual Input*: The native resolution scheme in Qwen2-VL [36] inspire us to advance RS VLMs to accept inputs of arbitrary shapes. As shown in Fig. 3, let  $H$  and  $W$  indicate the height and width of a given RS image.  $L_{patch}$  is the patch length corresponding to each visual token from the vision encoder. They first calculate the tightest shape that can wrap the input image by

$$(\hat{H}, \hat{W}) = (\lceil H/L_{patch} \rceil \times L_{patch}, \lceil W/L_{patch} \rceil \times L_{patch}), \quad (4)$$

where the  $\lceil \cdot \rceil$  means the ceiling function. Then, they resize the image to  $(\hat{H}, \hat{W})$  so that it can be exactly processed by the visual patch embedding.

This strategy allows the model to ingest images of arbitrary input sizes, which is well-suited to the diverse RS data. However, we should still set a range with a minimum scale to ensure adequate visual signal and a maximum scale due to constrained training resources. We divide the image sizes with the two bounds into three parts: small, regular, and UHR large. The small images are enlarged to ensure that there are enough visual tokens for the decoder to understand. For the UHR image, we design a zoom-in chain, which is introduced in III-D.

2) *Scalable Bounding Boxes*: For grounded or detection VLMs, spatial localization is achieved by directly generating numerical coordinates within textual outputs, which are extracted through regular expressions during inference.

However, existing RS VLMs often suffer from a mismatch between the coordinate resolution and the input image resolution. For instance, GeoChat [17] processes images at a fixed resolution of  $504 \times 504$ , but its coordinate resolution is only  $100 \times 100$ , leading to a fivefold loss in localization precision and poor performance on small objects. Conversely, GeoGround [19] employs a  $336 \times 336$  input resolution but defines coordinates at a much higher  $1000 \times 1000$  scale, resulting

in more than half of the coordinate space being unused and excessive localization precision.

In this work, we adopt scalable bounding boxes, whose coordinate resolution dynamically aligns with the input image resolution, thereby avoiding both under- and over-precision issues. This design naturally adapts to varying input sizes and allows flexible control of inference cost depending on the required localization accuracy.

3) *Random Resizing*: To ensure robust performance across varying input image sizes, we applied dynamic scale augmentation during training. For each task, input images were randomly rescaled to different resolutions. In grounding and detection tasks, the corresponding bounding boxes were synchronously scaled to maintain spatial consistency. We observed that this scale-based augmentation significantly improved the model's robustness to input-size variation. Moreover, the model trained under such conditions exhibited enhanced performance when performing high-resolution inference. This also enables a practical inference-time strategy, allowing users to adjust image resolution according to task requirements and computational constraints.

#### D. Zoom-in Chain for UHR RS Images

Previous works on understanding UHR RS images, such as LRS-VQA [21] and GeoLLaVA-8k [22], primarily focus on addressing the issue of excessive image tokens through visual token pruning. Although this approach has proven effective and computationally efficient, it typically requires additional training and is not well-suited for joint training with tasks using standard image resolutions.

We observed that when humans analyze UHR RS images on electronic devices, their workflow typically involves first scanning the entire image to identify regions of interest, then zooming into these regions before performing the actual task. Inspired by this workflow, we designed the Zoom-in Chain strategy for RS VLMs, as illustrated below:

$$\begin{aligned} \text{User : } & <\text{Prompt}> + \mathbf{I}_q + <\text{Question}> \\ \text{Assistant : } & [\mathbf{x}_1, \mathbf{y}_1, \mathbf{x}_2, \mathbf{y}_2] \\ \text{User : } & \text{Zoom-in}(\mathbf{I}_q, [\mathbf{x}_1, \mathbf{y}_1, \mathbf{x}_2, \mathbf{y}_2]) \\ \text{Assistant : } & <\text{Ground Truth}> \end{aligned} \quad (5)$$

The blue portions indicate the training labels, while the others are ignored for loss. Specifically, given a UHR RS image, we first downsample the image for initial processing. The model is prompted with instructions to predict the RoI, which is then cropped and fed into the model in native resolution. The final answer is obtained from both the initial and the new inputs, effectively mimicking the human zoom-in workflow for improved localization and task performance.

To enable the model to learn zoom-in capabilities during training, we construct a specialized instruction tuning dataset for UHR RS image perception, named LRS-VQA-Zoom. The data pipeline is initiated by collecting three public, large-scale UHR RS datasets: DOTA1.0 [48], GLH-Bridge [49], and STAR [50].

The methodology for generating the LRS-VQA-Zoom is extended from the pipeline in LRS-VQA [21]. The final

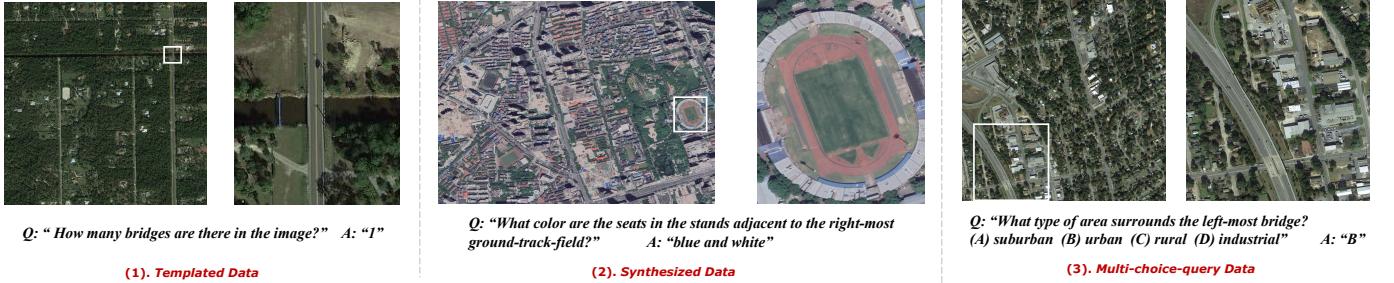


Fig. 4. Examples of the three types of annotated data in the proposed LRS-VQA-Zoom.

training corpus, totaling 302k samples, compromise three distinct subsets: 60k open-ended samples generated via rule-based templates, 159k open-ended samples synthesized using GPT-4V, and 83k samples in multi-choice-query form. Fig. 4 exhibits the examples from each subsets.

1) *Template-Generated Data (60k)*: This subset focuses on two open-ended question categories: counting and comparison. For the counting data, the UHR image is first divided into a  $3 \times 3$  grid (nine regions). Depending on the density of the target category, questions are formulated to query either the total count across the entire image or the count within a specific region. For the comparison data, these tasks involve comparing the relative quantities of two different object categories. For all samples in this subset, the absolute coordinates of the corresponding bounding boxes are preserved in the training data.

2) *GPT-4V-Synthesized Data (159k)*: This subset is designed to introduce greater question diversity. First, we filter the original object detection labels to identify unique target instances, which serve as “unique references”. Subsequently, the “coarse region” around each unique reference is cropped by applying a predefined padding margin. The dimensions of these coarse regions are suitable for processing by the GPT-4V model. We then prompt GPT-4V to generate diverse question-answer pairs based on these cropped regions. This process yields a rich variety of question types, including queries related to color, category, shape, status, spatial reasoning, and scene context (e.g., rural/urban). In this part of data, the coordinates of the horizontal bounding box defining the coarse region are recorded.

3) *Multi-choice-query Data (83k)*: To enhance the model’s proficiency with mainstream evaluation formats (i.e., multiple-choice query (MCQ)) and to further diversify the training data, we converted a subset of 83k open-ended question answering samples into an MCQ format using an automated pipeline centered around the large language models. For each question-answer pair, excluding simple binary (yes/no) queries, we prompted the GPT-4 to generate three plausible but incorrect “distractors” and return them alongside the original correct answer in a structured JSON format. This output was then systematically validated to ensure it contained four unique options. Finally, to prevent positional bias, the options were randomly shuffled, and the sample was formatted to include the question, four choices prefixed with letters (A, B, C, D), and the letter corresponding to the ground truth answer.

#### E. Auto-regressive Aerial Detection

In this paper, we investigate multi-class oriented aerial object detection. To enable the RS VLM to perform dense detection in aerial images, we propose an auto-regressive detection paradigm, representing detection outputs directly in textual form, as illustrated in the right part of Fig. 2. Specifically, we propose a normalization procedure for model responses and a novel evaluation metric to facilitate fair comparisons between the RS VLM detectors and conventional detectors.

1) *Response Normalization*: In the aerial object detection task, each object is represented by its class label and an 8-parameter quadrilateral bounding box  $\mathbf{o} = (n_{\mathbf{o}}, x_{1\mathbf{o}}, y_{1\mathbf{o}}, x_{2\mathbf{o}}, y_{2\mathbf{o}}, x_{3\mathbf{o}}, y_{3\mathbf{o}}, x_{4\mathbf{o}}, y_{4\mathbf{o}})$ , where  $(x_{i\mathbf{o}}, y_{i\mathbf{o}})$  denote the coordinates of the polygon vertices in clockwise order. The vertex with the smallest vertical coordinate is designated as the starting point. The class label  $n_{\mathbf{o}}$  corresponds to one of the  $c$  predefined categories  $\{C_1, C_2, \dots, C_c\}$ .

To standardize detection annotations, a consistent template is employed to ensure both uniqueness and order. For each input image, the model outputs detected objects in a structured sequence. Specifically, detection results are first grouped by category and sorted alphabetically by category name. Within each category, the bounding boxes are further ordered according to the position of their designated starting vertex.

During our extension of LMMRotate [20], we observed a subtle yet important issue. In the LMMRotate, images without any objects were removed from the training set to improve efficiency, following common practices in conventional aerial detector. However, this approach can be detrimental when training a VLM, as encountering object-free images during inference often leads the model to hallucinate, producing false positive detections. To address this, RSCoVLM retains images without objects in the training process and explicitly trains the model to output “There is none.” for such cases, thereby mitigating hallucinations and improving detection reliability.

Our VLM is capable of detecting multiple object categories within an aerial image, with both category labels and bounding box coordinates included in its output. During inference, detection results can be retrieved directly from the model response using straightforward regular expression parsing. Furthermore, unlike most traditional detectors that require post-processing procedures such as non-maximum suppression (NMS) to address overlapping or redundant detections, the VLM inherently avoids these issues.

2) *Evaluation Metrics*: In conventional aerial detection tasks, mean average precision (mAP) is widely employed as

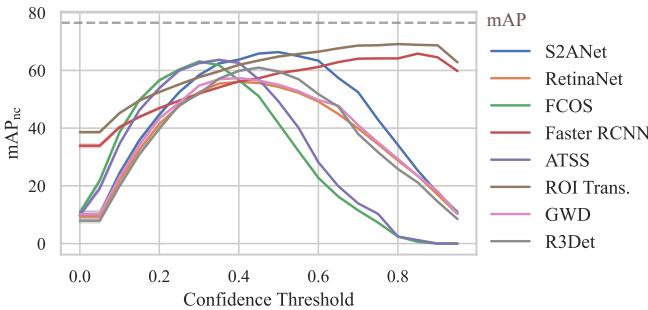


Fig. 5. The impact of confidence scores on  $mAP_{nc}$  with error bands. The colored lines record the variation trends of  $mAP_{nc}$  for the popular conventional detector on DOTA-v1.0 [48] (trained and evaluated on both the ‘train’ split and the ‘validation’ split) dataset under different confidence thresholds.

the evaluation metric, requiring bounding boxes, class labels, and confidence scores for all detected objects. However, as discussed earlier, our model responses only include object categories and their corresponding spatial coordinates. We observed that the reliability of confidence estimation considerably influences the final mAP, which poses an inherent limitation for VLMs.

Figure 5 illustrates the influence of confidence scores on performance. The curves depict mAP values obtained when the confidence terms are replaced by either fixed or random constants, denoted as mAP with no confidence scores, i.e.,  $mAP_{nc}$ . Despite the application of NMS in conventional detectors, their direct  $mAP_{nc}$  remains low due to numerous low-confidence false positives. To provide a fairer comparison, we perform threshold-based filtering to maximize  $mAP_{nc}$  for conventional detectors. The figure clearly shows that incorporating confidence scores leads to a noticeable increase in mAP.

Instead of introducing an additional mechanism to estimate confidence for MLM-based detectors, we argue that confidence should not be a prerequisite when evaluating or comparing detection performance between MLMs and conventional detectors. Detection annotations and outputs inherently consist of class labels and bounding boxes, while confidence scores are auxiliary byproducts generated during inference. They may facilitate post-processing but are not indispensable for evaluating model accuracy. Therefore, we advocate employing confidence-independent metrics such as mean F1-score ( $mF_1$ ) and  $mAP_{nc}$  for a more equitable evaluation.

To further verify the robustness of  $mAP_{nc}$ , we repeated each evaluation eleven times by substituting confidence values with ten random and one constant setting. As shown in Fig. 5, the resulting standard deviations remain below 0.5%, confirming its stability as a metric.

Finally, for benchmarks such as DOTA [48] and FAIR1M [51], where public test sets are unavailable and online evaluation servers rely solely on mAP, we recommend adopting  $mAP_{nc}$  as the primary evaluation metric to ensure consistent and fair assessment across different model types.

#### IV. EXPERIMENT

In this section, the RSCoVLM is evaluated on benchmarks across various tasks, demonstrating the promising multi-task

capabilities. We firstly provide detailed implementation specifications to facilitate reproducibility. Then, we compare our model with state-of-the-art methods on various RS understanding and perception tasks with different input resolutions.

Fig. 6 presents the demonstration of RSCoVLM’s capabilities on several commonly used tasks. Notably, all tasks are accomplished using a single RSCoVLM model, demonstrating its impressive multi-task capability.

##### A. Reproducibility Details

We use Qwen2.5-VL-7B-Instruct [37] as the foundation model of RSCoVLM. The model is optimized with AdamW, employing a weight decay of 0.1. We train the full model with a base learning rate of  $2 \times 10^{-6}$ , following a cosine learning rate schedule with a linear warmup over the first 5% of training steps. The total batch size is set to 32, and the maximum sequence length is 6,144 tokens. The input images are constrained to resolutions between  $224 \times 224$  and  $1,008 \times 1,008$  pixels.

We have released the codebase on the GitHub repository and uploaded the whole well-collected data folder and model weights to the HuggingFace repository. The codebase is implemented concisely, leveraging resource-efficient and effective training techniques. To save GPU memory, we adopt DeepSpeed-ZeRO-Stage-1 [58] and gradient checkpointing. For improved computational efficiency, we utilize BFloat16 precision and Flash-Attention-2 [59] during both training and evaluation. Additionally, Liger Kernel [60] is employed to accelerate training, and vLLM [61] is used for faster inferencing. All experiments are conducted on VolcEngine high-performance computing clusters equipped with NVIDIA A800 GPUs. We’ll maintain the repositories and update the latest code, model and data in our future research progress.

##### B. Evaluation on Large RS Imagery

1) *Benchmark and Metric:* The LRS-VQA [21] is the latest visual question answering benchmark for large RS images. It features 7,333 question-answer pairs across 8 categories, including count, color, category, shape, status, reasoning, rural/urban classification, and target background. The images in this benchmark reach up to 27,328 pixels in length and have an average size of  $7,099 \times 6,329$  pixels.

There are three subsets, corresponding to three data sources: FAIR1M [51], GLH-Bridge [49], and STAR [50]. The official scoring implementation first calculates accuracy for each source and task, and then computes average accuracy (AA) across tasks for each source. The AAs for each subset are reported.

2) *Results:* The results are presented in Table I. The max pixels numbers of each models are also provided. It can be seen that the average pixel number of LRS-VQA (about 45 million) has been larger than the largest pixels uplimit (16.8 million for Qwen3-VL [37]).

As shown in the table, the proposed Zoom-in Chain approach substantially enhances the model’s performance, achieving an overall improvement of 35% compared to the baseline that inference solely.

**The proposed RSCoVLM performs multiple tasks independently and simultaneously**

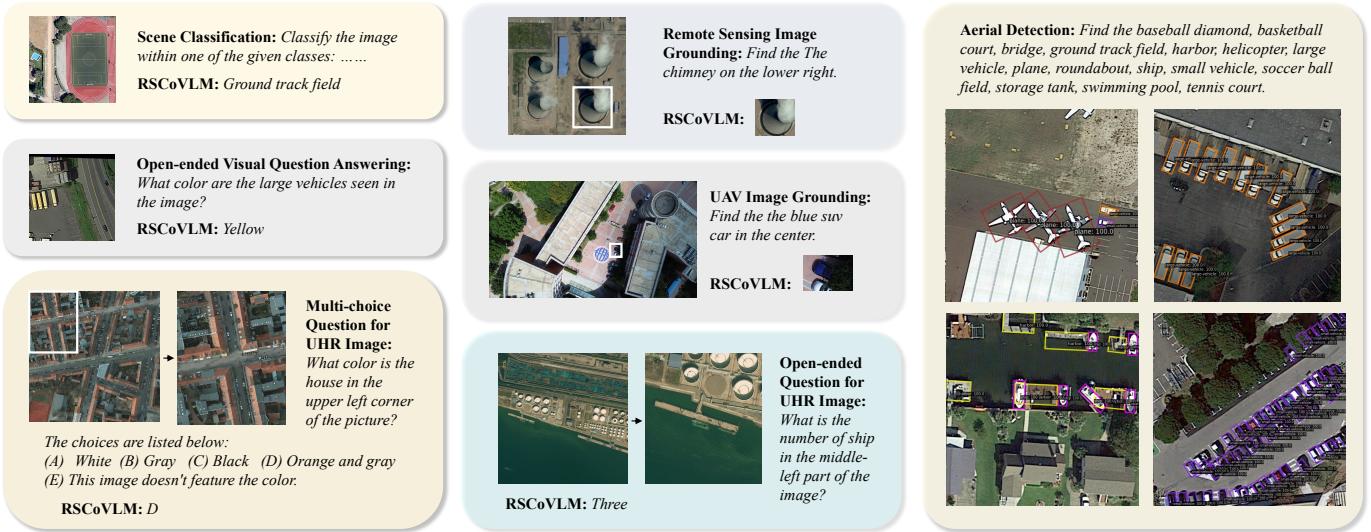


Fig. 6. Demonstration of RSCoVLM’s capabilities on several commonly used tasks, including scene classification, open-ended and multiple-choice question answering for regular and UHR images, visual grounding in aerial and UAV images, and aerial object detection. In particular, the visualized results of aerial detection are especially impressive.

TABLE I  
COMPARISON RESULTS OF STATE-OF-THE-ART VISION-LANGUAGE MODELS AND OUR MODEL ON THE LRS-VQA BENCHMARK

Method	Model Size	Max pixels	LRS-FAIR	LRS-Bridge	LRS-STAR	Avg. Acc
LLaVA-1.5 [52]	7B	0.1M	18.76	30.70	22.63	24.03
LLaVA-UHD-v2 [53]	7B	0.7M	22.82	32.57	26.08	27.16
Qwen2-VL [36]	7B	11.1M	23.80	38.12	27.87	29.93
Qwen2.5-VL [37]	7B	12.8M	19.66	35.82	26.70	27.39
Qwen3-VL [37]	8B	16.8M	<b>27.98</b>	<b>38.56</b>	<b>32.04</b>	<b>32.86</b>
	A3B-30B	16.8M	<b>27.63</b>	38.81	30.54	32.33
InternVL2.5-MPO [54]	8B	2.4M	24.95	34.59	25.14	28.23
InternVL3 [55]	8B	2.4M	22.49	38.09	26.36	28.98
InternVL3.5 [56]	8B	2.4M	25.14	35.50	26.86	29.17
	A3B-30B	2.4M	16.83	37.05	22.15	25.34
Mimo-VL [57]	7B	12.8M	16.51	20.04	27.11	21.22
GeoChat [17]	7B	0.3M	20.18	24.54	13.75	19.49
LLaVA-1.5 + SFT. on LRS-VQA [21]	7B	0.1M	22.97	36.89	27.48	29.11
LLaVA-Next + SFT. on LRS-VQA [21]	7B	2.8M	21.85	38.24	26.67	28.92
<b>RSCoVLM</b>	<b>7B</b>	<b>1.0M</b>	27.37	<b>42.42</b>	<b>31.77</b>	<b>33.85</b>
+ Zoom-in Chain			<b>42.42</b>	<b>49.56</b>	<b>45.15</b>	<b>45.71</b>

Furthermore, our model demonstrates stronger foundational capabilities than other competing models, approaching the performance of the leading Qwen3-VL-8B [37], while utilizing a slightly smaller parameter count and a significantly lower maximum input resolution. Our model also outperforms other RS foundation models, including GeoChat and the officially fine-tuned LLaVA-Next model for LRS-VQA [21].

### C. Evaluation on Visual Grounding

1) *Benchmark and Metric:* We follow GeoGround [19] for visual grounding evaluation because of its strong emphasis on comprehensiveness, fairness, and transparency. The evaluation incorporates the validation and test sets of DIOR-RSVG

and RSVG [62], the visual grounding portions of GeoChat-Bench [17] and VRSBench [63], as well as AVVG benchmark [19] for images captured by unmanned aerial vehicle. The evaluation details have strictly aligned with GeoGround. We directly adopted the splits and annotations provided by GeoGround for all benchmarks [19].

We follow common practice to utilize Acc@0.5 as the evaluation metric, which regards the predicting that has an Intersection over Union (IoU) greater than 0.5 with the ground truth as a successful localizing.

2) *Results:* Table II presents the results, along with the corresponding input sizes for each model. The input resolutions of existing RS VLMs, including GeoChat [17], LHRSS-Bot [18], VHM [16], and GeoGround [19], are fixed and

TABLE II  
COMPARISON RESULTS OF STATE-OF-THE-ART VISION-LANGUAGE MODELS AND OUR MODEL ON VISUAL GROUNDING BENCHMARKS

Method	Input Size	DIOR-RSVG		RSVG		GeoChat-VG	VRSBench-VG	AVVG	Avg. Acc.
		val	test	val	test				
Qwen-VL-Chat [36]	448 × 448	32.01	32.22	4.66	2.04	35.36	31.07	0.31	19.66
GeoChat [17]	504 × 504	23.35	24.05	3.08	2.04	22.74	11.52	0.28	12.44
LHRS-Bot [18]	224 × 224	17.04	17.59	0.95	1.56	3.25	1.19	0.00	5.94
VHM [16]	336 × 336	-	48.04	-	-	-	-	-	-
Qwen2.5-VL [37]	Dynamic	43.64	45.26	19.73	21.27	42.99	44.50	7.64	32.15
Qwen-VL + SFT. on refGeo [19]	448 × 448	58.65	58.76	12.99	10.59	41.75	47.38	9.53	34.24
GeoChat + SFT. on refGeo [19]	504 × 504	60.27	61.96	16.32	14.67	56.99	51.36	11.52	39.01
LLaVA-1.5-7B + SFT. on refGeo [19]	336 × 336	64.46	65.98	19.98	20.95	63.76	57.17	15.05	43.91
GeoGround [19]	336 × 336	<b>77.18</b>	<b>77.73</b>	<b>27.64</b>	<b>26.65</b>	<b>70.24</b>	66.04	<b>21.58</b>	<b>52.44</b>
<b>RSCoVLM</b>	Dynamic	<b>83.56</b>	<b>84.55</b>	<b>54.04</b>	<b>53.79</b>	<b>76.39</b>	<b>79.73</b>	<b>29.40</b>	<b>65.92</b>
+ Min Size	224 × 224	66.56	67.64	21.23	20.70	21.43	<b>67.50</b>	0.85	37.99
+ Small Size	336 × 336	<b>75.22</b>	<b>75.86</b>	<b>34.72</b>	<b>35.79</b>	<b>70.17</b>	<b>75.79</b>	<b>25.10</b>	<b>56.09</b>

typically smaller than 512×512. In contrast, only general-purpose VLMs such as Qwen2.5-VL [37] and MiMo-VL [57] support dynamic input resolution, enabling flexible adaptation to varying input sizes.

Our model demonstrates substantially superior performance across all benchmarks. It surpasses the previously best-performing visual-language model specialized for RS grounding, GeoGround, by approximately 25.7%, and outperforms all baselines that were supervised-finetuned on refGeo.

We further conducted experiments using fixed low-resolution inputs to intentionally weaken our model’s performance. Even at the minimal input size of 224×224, our model maintains strong capability; however, such a small resolution severely limits image clarity, causing small objects to occupy only a few pixels and become indistinguishable. In particular, performance on AVVG drops sharply, indicating that a 224×224 resolution is highly impractical for RS grounding. When evaluated at 336×336, which aligns with the input size of other comparison methods, our model still achieves state-of-the-art results.

We attribute this performance advantage to three primary factors. First, the support for dynamic input resolution allows the model to perform inference at native resolution without downsampling, preserving visual detail. Second, the multi-resolution augmentation strategy employed during training enables the model to generalize effectively across diverse resolutions and computational budgets. Finally, auxiliary localization-related tasks, such as object detection and zoom-in refinement, further strengthen the model’s grounding ability and robustness.

#### D. Evaluation on Object Detection

1) *Benchmark, Metric, and Comparison Setting:* We selected the most widely-used aerial image object detection benchmark, DOTA-v1.0 [48], for our evaluation. The whole DOTA-v1.0 dataset comprises 2,806 high-resolution aerial images and 188,282 object instances across 15 common categories. The proportions of testing set is 1/3. These images were collected from multiple sensors and platforms, and each

instance is annotated with a 8 degrees-of-freedom oriented bounding box, capturing the wide variations in object scale, shape, and orientation typical of aerial imagery.

We adopt the Average Precision with no confidence ( $AP_{nc}$ ) and report three specific variants:  $AP_{nc50}$  (IoU threshold is 0.50),  $AP_{nc75}$  (IoU threshold is 0.75), and  $AP_{nc50:95}$  (the average  $AP_{nc}$  computed over IoU thresholds from 0.50 to 0.95 at increments of 0.05). The evaluation is based on the standard MMRotate [77] evaluation procedure. And the splitting length is set 512 with an overlap of 100.

The conventional object detection baselines are trained using the latest MMRotate [77], and the details necessary for reproducibility are also provided in the released code. We obtain a reasonable  $AP_{nc}$  for comparison methods using the following procedure: We first select a threshold for confidence scores to filter out low-score predictions, and then randomize (or, set to 1) the remaining prediction scores. The AP computed under this condition is denoted as  $AP_{nc}$  of the conventional detector. To determine an appropriate threshold for each detector, we evaluate  $AP_{nc}$  on the validation set by varying the confidence threshold from 0.00 to 0.95 in increments of 0.05, and select the threshold that yields the highest  $AP_{nc}$  for subsequent evaluation on the test set.

2) *Results:* We compare our model with state-of-the-art RS object detection methods. Our multi-task model achieves detection performance comparable to conventional detectors, even though it is not specifically optimized for the single dataset as the comparison methods are. When trained solely on object detection data, denoted as RSCoVLM-det, the model exhibits further improvement and even surpasses half of the conventional methods. It is a remarkable achievement for RS vision-language models.

Thanks to the dynamic resolution strategy, our model can further enhance detection performance by maximizing the inference scale, referred to as the “Max Mode.” Specifically, each input image is upsampled to the model’s upper input limit of 1008×1008, and the outputs are then downsampled back to the original scale for evaluation. We observe a substantial increase in overall  $AP_{nc}$ , although certain categories such as plane (PL) and bridge (BD) experience minor degradation.

TABLE III  
COMPARISON RESULTS OF STATE-OF-THE-ART AERIAL DETECTORS AND OUR MODEL ON DOTA-v1.0 BENCHMARK

Method	score	PL	BD	BR	GTF	SV	LV	SH	TC	BC	ST	SBF	RA	HA	SP	HC	AP <sub>nc50</sub>	AP <sub>nc75</sub>	AP <sub>nc50:95</sub>
GWD [64]	0.40	72.07	58.95	24.26	35.92	63.37	52.24	66.63	86.85	61.54	59.15	23.45	45.84	36.10	49.50	26.08	50.80	28.67	28.98
R3Det [65]	0.45	73.32	59.51	31.59	43.37	<b>64.94</b>	63.37	<b>75.61</b>	<b>89.35</b>	63.84	66.95	<b>34.99</b>	45.60	46.54	50.17	14.48	54.91	29.21	30.08
ATSS [66]	0.35	72.98	60.67	25.82	42.91	<b>65.23</b>	<b>65.32</b>	75.22	<b>89.78</b>	<b>71.61</b>	<b>70.12</b>	28.04	43.19	47.79	58.28	28.60	56.37	<b>34.62</b>	<b>33.05</b>
Faster RCNN [67]	0.85	73.49	<b>67.37</b>	<b>32.50</b>	43.19	62.92	63.13	73.95	88.79	<b>73.77</b>	66.33	25.23	<b>48.89</b>	<b>53.03</b>	56.94	31.17	<b>57.38</b>	<b>32.98</b>	<b>32.96</b>
FCOS [68]	0.30	72.09	56.32	32.02	27.79	64.28	63.83	<b>75.75</b>	89.21	68.11	<b>67.82</b>	27.30	37.15	46.64	<b>58.98</b>	19.53	53.79	32.13	31.49
CSL [69]	0.40	71.57	53.57	19.82	35.90	64.04	44.96	66.35	87.54	61.33	59.26	<b>29.27</b>	39.56	36.51	49.80	17.38	49.12	29.03	28.72
S2A-Net [70]	0.50	72.96	61.96	<b>36.00</b>	<b>45.99</b>	<b>66.24</b>	<b>65.61</b>	<b>77.08</b>	<b>89.34</b>	<b>73.27</b>	<b>69.25</b>	<b>31.78</b>	<b>46.64</b>	<b>55.02</b>	52.02	35.40	<b>58.58</b>	29.53	31.98
<b>RSCoVLM</b>		<b>77.15</b>	<b>64.86</b>	23.90	<b>45.34</b>	44.87	38.96	57.64	87.22	57.73	49.42	23.31	51.87	37.01	54.92	<b>54.91</b>	51.27	25.75	27.60
+ Max Mode		<b>73.95</b>	63.01	27.84	40.41	56.86	55.37	71.00	89.12	61.69	64.95	19.54	41.91	44.21	55.01	<b>52.27</b>	54.48	31.04	31.38
<b>RSCoVLM-det</b>		<b>73.52</b>	<b>64.68</b>	26.89	<b>47.18</b>	52.57	52.71	59.33	89.16	63.17	61.43	18.91	45.96	47.62	<b>59.19</b>	<b>70.24</b>	55.50	30.78	31.75
+ Max Mode		69.04	64.44	<b>33.32</b>	44.67	56.21	<b>66.47</b>	73.71	87.59	61.38	63.95	22.41	<b>46.63</b>	<b>47.90</b>	<b>59.82</b>	50.82	<b>56.56</b>	<b>33.88</b>	<b>33.66</b>

TABLE IV  
COMPARISON RESULTS OF STATE-OF-THE-ART VISION-LANGUAGE MODELS AND OUR MODEL ON FIVE SCENE CLASSIFICATION BENCHMARKS

Method	Model Size	AID	UCMerced	METER-ML	NWPU-RESISC45	WHU-RS19
MiniGPTv2 [71]	7B	-   32.96	-	14.29	28.15	64.80
LLaVA-1.5 [52]	7B	-   31.10	-	21.73	34.96	54.55
Qwen-VL-Chat [36]	7B	-   55.30	-	38.77	42.73	72.25
Qwen2.5-VL [37]	7B	63.63   62.73	70.90	56.64	64.98	76.20
Qwen3-VL [37]	8B	70.84   66.67	79.90	60.88	68.86	87.80
	A3B-30B	71.75   68.87	80.19	64.07	70.22	87.70
InternVL2.5-MPO [54]	8B	69.38   64.23	62.90	55.04	59.21	80.20
InternVL3 [55]	8B	67.78   63.40	67.29	59.65	64.32	86.40
InternVL3.5 [56]	8B	77.03   75.00	83.43	51.33	92.57	91.70
	A3B-30B	82.45   79.17	<b>86.00</b>	46.19	<b>98.38</b>	<b>97.10</b>
Mimo-VL [57]	7B	66.13   67.20	69.14	54.51	64.35	86.10
LHRSBot [18]	7B	-   <b>91.26</b>	-	69.81	83.94	93.17
GeoChat [17]	7B	72.00   -	84.40	-	-	-
TEOChat [72]	7B	80.90   -	<b>86.30</b>	-	-	-
LHRS-Bot-Nova [73]	7B	<b>83.06</b>   -	-	<b>72.74</b>	83.97	<b>96.20</b>
SkysenseGPT [74]	7B	<b>88.16</b>   -	-	40.00	90.06	95.50
VHM [16]	7B	-   <b>91.70</b>	-	<b>72.74</b>	<b>94.54</b>	95.80
ScoreRS [75]	7B	-   85.90	-	<b>74.42</b>	91.59	<b>96.30</b>
<b>RSCoVLM</b>	<b>7B</b>	<b>88.44</b>   <b>94.30</b>	<b>94.52</b>	<b>75.93</b>	<b>98.25</b>	95.80

The enhanced RSCoVLM-det even outperforms all competing approaches, while the conventional detectors are trained and evaluated at fixed resolutions without such a feature of test-time augmentation.

To the best of our knowledge, only two existing RS VLMs, LMMRotate [20] (our conference version) and Falcon [78], are capable of performing aerial object detection effectively. However, their common foundation model, Florence-2 [79], employs a fixed and relatively large input size of  $1024 \times 1024$ , which already exceeds the input limit of RSCoVLM. Moreover, LMMRotate is trained specifically for detection, while Falcon performs well only on its training set and does not report test results. In addition, Falcon requires multiple inferences per image, making separate predictions for each category, which results in extremely high computational cost. Therefore, we consider RSCoVLM to be the only vision-

language model capable of performing multiple tasks while achieving detection performance that is fairly comparable to specialized object detection models, currently.

#### E. Evaluation on Scene Classification

1) *Benchmark and Metric:* We evaluate our model on five standard remote-sensing scene-classification benchmarks. The AID [80] dataset comprises approximately 10,000 images of size  $600 \times 600$  pixels across 30 classes. The UCMerced [81] dataset consists of 2,100 images of size  $256 \times 256$  pixels covering 21 classes. The NWPU-RESISC45 [82] dataset contains 31,500 images of size  $256 \times 256$  pixels across 45 classes, with large variation in resolution and scene complexity. The WHU-RS19 [83] dataset includes around 1,000 high-resolution patches of size  $600 \times 600$  pixels spanning 19 classes. The METER-ML [84] benchmark offers a large-scale

TABLE V  
COMPARISON RESULTS OF STATE-OF-THE-ART VISION-LANGUAGE MODELS AND OUR MODEL ON TWO VQA BENCHMARKS

Method	RSVQA Benchmark						VRSBench VQA
	HR-Comp.	HR-Pres.	LR-Comp.	LR-Pres.	LR-R-U	Avg.	
LLaVA-1.5 [52]	67.30	69.80	68.20	55.50	59.00	63.96	-
LLaVA-1.6 [75]	68.60	64.40	64.32	56.84	61.00	63.03	-
Qwen2-VL [36]	75.60	63.30	75.47	62.00	73.00	69.87	-
Qwen2.5-VL [37]	75.28	67.30	73.86	64.67	66.00	69.42	51.21
Qwen3-VL [37]	81.00	78.10	70.32	56.42	72.00	71.57	54.75
InternVL-2.5 [76]	75.50	65.80	71.16	66.21	72.00	70.13	47.20
InternVL3 [55]	74.15	62.35	73.06	66.23	74.00	69.96	50.68
InternVL3.5 [56]	80.14	53.80	92.00	91.26	96.00	82.64	53.74
Mimo-VL [57]	66.00	77.80	74.42	59.98	65.00	68.64	48.37
LHRS-Bot-Nova [73]	89.30	87.60	88.11	83.89	79.00	85.58	-
GeoChat [17]	83.30	59.10	90.52	90.63	97.00	84.11	40.80
VHM [16]	83.30	68.30	90.11	89.89	87.00	83.72	-
<b>RSCoVLM</b>	82.60	68.50	93.16	92.18	94.00	<b>86.09</b>	<b>58.08</b>

multi-sensor setup with varied image sizes for extended generalisation evaluation. Together these benchmarks allow a robust assessment of our model’s generalisation across dataset scale, class-set size, imaging conditions and spatial resolutions.

We report overall accuracy of the test set for each benchmark. For METER-ML, NWPU-RESISC45, and WHU-RS19, we adopt the test set splits defined by VHM [16]. For UCMerced, we follow the split defined by GeoChat [17]. For AID, we present results using both the VHM and GeoChat splits to facilitate fair comparison.

2) *Results:* Table IV presents the comparative results across the five scene classification benchmarks. The compared methods include classical VLM baselines (MiniG-PTv2 [71] and LLaVA-1.5 [52]), leading open-source VLMs (the QwenVL [37], InternVL [76], and MiMo-VL [57] series), and latest RS VLMs (GeoChat [17], TEOChat [72], LHRS-Bot-Nova [73], SkysenseGPT [74], VHM [16], and ScoreRS [75]). As shown, our model consistently surpasses all compared approaches across all benchmarks.

#### F. Evaluation on Visual Question Answering

1) *Benchmark and Metric:* We evaluate our model’s visual question answering capability using two established benchmarks in the RS domain, including the RSVQA benchmark [85] and the VQA portion of VRSBench [63]. The RSVQA compromise two subsets of image-question-answer triplet derived from high-resolution (HR) orthorectified imagery and low-resolution (LR) RS data, enabling evaluation of model reasoning across spatial scales. The VRSBench dataset is a large-scale vision-language benchmark for RS image understanding that comprises 37,408 question-answer pairs in test set, supporting a broad range of understanding instructions. The standard question answering accuracy is used as the metric.

2) *Results:* Table V presents the results of visual question answering, demonstrating the strong understanding and conversational capabilities of our model. Our approach surpasses all open-source VLMs (including the LLaVA, Qwen,

InternVL, and MiMo-VL series) as well as RS VLMs (GeoChat [17], LHRS-Bot-Nova [73], and VHM [16]) across the two benchmarks. In the zero-shot question answering evaluation on VRSBench-VQA [63], our model outperforms the latest general-purpose models, showing superior generalization ability on RS image question answering.

#### V. CONCLUSION

In this paper, we introduce RSCoVLM, the latest generation of versatile vision language model. We carefully curated RS data, detailing the processes of data collection, offline integration, and online loading with adaptive weighting. To handle the wide range of image resolutions in RS images, we developed a dynamic-resolution strategy and proposed the Zoom-in Chain mechanism with the LRS-VQA-Zoom dataset for ultra-high-resolution images. Moreover, we improved the model’s object detection capabilities and designed a fair evaluation protocol for comparison with conventional methods. Comprehensive experiments show that RSCoVLM consistently delivers state-of-the-art results across multiple tasks, surpassing previous RS VLMs and matching task-specific expert models. By releasing all code, models, and datasets, we aim to enable reproducibility and foster progress toward general-purpose remote sensing models.

#### REFERENCES

- [1] L. Zhang and L. Zhang, “Artificial intelligence for remote sensing data analysis: A review of challenges and opportunities,” *IEEE Geosci. Remote Sens. Magaz.*, vol. 10, no. 2, pp. 270–294, 2022.
- [2] Y. Zhou, L. Feng, Y. Ke *et al.*, “Towards vision-language geo-foundation models: A survey,” *arXiv preprint arXiv:2406.09385*, 2024.
- [3] Y. Zhang and Q. Yang, “A survey on multi-task learning,” *IEEE Trans. on Know. and Data Engin.*, vol. 34, no. 12, pp. 5586–5609, 2022.
- [4] Q. Li, Y. Chen, X. He, and L. Huang, “Co-training transformer for remote sensing image classification, segmentation, and detection,” *IEEE Transactions on Geoscience and Remote Sensing*, vol. 62, pp. 1–18, 2024.
- [5] A. Vaswani, N. Shazeer, N. Parmar *et al.*, “Attention is all you need,” *NIPS*, vol. 30, 2017.
- [6] A. Dosovitskiy, L. Beyer, A. Kolesnikov *et al.*, “An image is worth 16x16 words: Transformers for image recognition at scale,” in *ICLR*, 2021.

- [7] Q. Li, Y. Chen, and Y. Zeng, "Transformer with transfer cnn for remote-sensing-image object detection," *Remote Sens.*, vol. 14, no. 4, p. 984, 2022.
- [8] J. Han, K. Gong, Y. Zhang *et al.*, "Onellm: One framework to align all modalities with language," 2023.
- [9] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever, "Learning transferable visual models from natural language supervision," in *ICML*, 2021.
- [10] Q. Li, Z. Chen, W. Wang *et al.*, "Omnicorpus: A unified multimodal corpus of 10 billion-level images interleaved with text," in *The Thirteenth International Conference on Learning Representations*, 2025.
- [11] J.-B. Alayrac, J. Donahue, P. Luc *et al.*, "Flamingo: a visual language model for few-shot learning," in *Advances in Neural Information Processing Systems*, vol. 35. Curran Associates, Inc., 2022, pp. 23 716–23 736.
- [12] Z. Chen, J. Wu, W. Wang *et al.*, "Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 24 185–24 198.
- [13] H. Liu, C. Li, Q. Wu, and Y. J. Lee, "Visual instruction tuning," in *Advances in Neural Information Processing Systems*, A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, Eds., vol. 36. Curran Associates, Inc., 2023, pp. 34 892–34 916.
- [14] D. Wang, J. Zhang, M. Xu *et al.*, "Mtp: Advancing remote sensing foundation model via multi-task pretraining," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, pp. 1–24, 2024.
- [15] Y. Li, X. Li, Y. Li *et al.*, "Sm3det: A unified model for multi-modal remote sensing object detection," *arXiv preprint arXiv:2412.20665*, 2024.
- [16] C. Pang, X. Weng, J. Wu *et al.*, "Vhm: Versatile and honest vision language model for remote sensing image analysis," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 39, no. 6, 2025, pp. 6381–6388.
- [17] K. Kuckreja, M. S. Danish, M. Naseer *et al.*, "Geochat: Grounded large vision-language model for remote sensing," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2024, pp. 27 831–27 840.
- [18] D. Muhtar, Z. Li, F. Gu *et al.*, "Lhrs-bot: Empowering remote sensing with vgi-enhanced large multimodal language model," in *European Conference on Computer Vision*. Springer, 2024, pp. 440–457.
- [19] Y. Zhou, M. Lan, X. Li *et al.*, "Geoground: A unified large vision-language model for remote sensing visual grounding," 2024.
- [20] Q. Li, Y. Chen, X. Shu *et al.*, "A simple aerial detection baseline of multimodal language models," *arXiv preprint arXiv:2501.09720*, 2025.
- [21] J. Luo, Y. Zhang, X. Yang *et al.*, "When large vision-language model meets large remote sensing imagery: Coarse-to-fine text-guided token pruning," *arXiv preprint arXiv:2503.07588*, 2025.
- [22] F. Wang, M. Chen, Y. Li *et al.*, "Geollava-8k: Scaling remote-sensing multimodal large language models to 8k resolution," *arXiv preprint arXiv:2505.21375*, 2025.
- [23] W. Zhang, M. Cai, T. Zhang *et al.*, "Earthgpt: A universal multimodal large language model for multisensor image comprehension in remote sensing domain," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 62, pp. 1–20, 2024.
- [24] J. Luo, Z. Pang, Y. Zhang *et al.*, "Skysensegt: A fine-grained instruction tuning dataset and model for remote sensing vision-language understanding," *arXiv preprint arXiv:2406.10100*, 2024.
- [25] J. Chen, H. Guo, K. Yi *et al.*, "Visualgpt: Data-efficient adaptation of pretrained language models for image captioning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2022, pp. 18 030–18 040.
- [26] J. Li, D. Li, S. Savarese, and S. Hoi, "BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models," in *Proceedings of the 40th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, A. Krause, E. Brunskill, K. Cho, B. Engelhardt, S. Sabato, and J. Scarlett, Eds., vol. 202. PMLR, 23–29 Jul 2023, pp. 19 730–19 742.
- [27] D. Zhu, J. Chen, X. Shen, X. Li, and M. Elhoseiny, "MiniGPT-4: Enhancing vision-language understanding with advanced large language models," in *The Twelfth International Conference on Learning Representations*, 2024.
- [28] W. Dai, J. Li, D. Li *et al.*, "Instructblip: Towards general-purpose vision-language models with instruction tuning," in *Advances in Neural Information Processing Systems*, A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, Eds., vol. 36. Curran Associates, Inc., 2023, pp. 49 250–49 267.
- [29] P. Gao, J. Han, R. Zhang *et al.*, "Llama-adapter v2: Parameter-efficient visual instruction model," 2023.
- [30] Z. Lin, C. Liu, R. Zhang *et al.*, "Sphinx: The joint mixing of weights, tasks, and visual embeddings for multi-modal large language models," 2023.
- [31] W. Wang, Z. Chen, X. Chen *et al.*, "Visionllm: Large language model is also an open-ended decoder for vision-centric tasks," in *Advances in Neural Information Processing Systems*, A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, Eds., vol. 36. Curran Associates, Inc., 2023, pp. 61 501–61 513.
- [32] J. Wu, M. Zhong, S. Xing *et al.*, "Visionllm v2: An end-to-end generalist multimodal large language model for hundreds of vision-language tasks," in *Advances in Neural Information Processing Systems*, A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang, Eds., vol. 37. Curran Associates, Inc., 2024, pp. 69 925–69 975.
- [33] X. Chen, J. Djolonga, P. Padlewski *et al.*, "Pali-x: On scaling up a multilingual vision and language model," 2023.
- [34] Z. Yue, Z. Lin, Y. Song *et al.*, "Mimo-vl technical report," 2025.
- [35] W. Wang, Q. Lv, W. Yu *et al.*, "Cogvlm: Visual expert for pretrained language models," in *Advances in Neural Information Processing Systems*, A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang, Eds., vol. 37. Curran Associates, Inc., 2024, pp. 121 475–121 499.
- [36] P. Wang, S. Bai, S. Tan *et al.*, "Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution," 2024.
- [37] S. Bai, K. Chen, X. Liu *et al.*, "Qwen2.5-vl technical report," 2025.
- [38] J. A. Irvin, E. R. Liu, J. C. Chen *et al.*, "Teochat: A large vision-language assistant for temporal earth observation data," *arXiv preprint arXiv:2410.06234*, 2024.
- [39] F. Cui and J. Jiang, "Mtscd-net: A network based on multi-task learning for semantic change detection of bitemporal remote sensing images," *International Journal of Applied Earth Observation and Geoinformation*, vol. 118, p. 103294, 2023.
- [40] Y. Wang, L. Zhao, Y. Hu, H. Dai, and Y. Zhang, "Multitask semantic change detection guided by spatiotemporal semantic interaction," *Scientific Reports*, vol. 15, no. 1, p. 16003, 2025.
- [41] Y. Niu, H. Guo, J. Lu, L. Ding, and D. Yu, "Smnet: Symmetric multi-task network for semantic change detection in remote sensing images based on cnn and transformer," *Remote Sensing*, vol. 15, no. 4, p. 949, 2023.
- [42] H. Lin, X. Wang, M. Li, D. Huang, and R. Wu, "A multi-task consistency enhancement network for semantic change detection in hr remote sensing images and application of non-agriculturalization," *Remote Sensing*, vol. 15, no. 21, p. 5106, 2023.
- [43] M. Lu, J. Liu, F. Wang, and Y. Xiang, "Multi-task learning of relative height estimation and semantic segmentation from single airborne rgb images," *Remote Sensing*, vol. 14, no. 14, p. 3450, 2022.
- [44] M. Carvalho, B. Le Saux, P. Trouvé-Peloux, F. Champagnat, and A. Almansa, "Multitask learning of height and semantics from aerial images," *IEEE Geoscience and Remote Sensing Letters*, vol. 17, no. 8, pp. 1391–1395, 2019.
- [45] F. Bastani, P. Wolters, R. Gupta, J. Ferdinando, and A. Kembhavi, "Satlaspretrain: A large-scale dataset for remote sensing image understanding," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2023, pp. 16 772–16 782.
- [46] S. Soni, A. Dudhane, H. Debary *et al.*, "Earthdial: Turning multi-sensory earth observations to interactive dialogues," in *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025, pp. 14 303–14 313.
- [47] B. Li, Y. Zhang, D. Guo *et al.*, "LLaVA-onevision: Easy visual task transfer," *Transactions on Machine Learning Research*, 2025.
- [48] G.-S. Xia, X. Bai, J. Ding *et al.*, "DOTA: A large-scale dataset for object detection in aerial images," in *CVPR*, June 2018.
- [49] Y. Li, J. Luo, Y. Zhang *et al.*, "Learning to holistically detect bridges from large-size vhr remote sensing imagery," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 11, pp. 7778–7796, 2024.
- [50] Y. Li, L. Wang, T. Wang *et al.*, "Star: A first-ever dataset and a large-scale benchmark for scene graph generation in large-size satellite imagery," *arXiv preprint arXiv:2406.09410*, 2024.
- [51] X. Sun, P. Wang, Z. Yan *et al.*, "Fair1m: A benchmark dataset for fine-grained object recognition in high-resolution remote sensing imagery," *ISPRS J. Photogram. Remote Sens.*, vol. 184, pp. 116–130, 2022.

- [52] H. Liu, C. Li, Y. Li, and Y. J. Lee, “Improved baselines with visual instruction tuning,” 2023.
- [53] Y. Zhang, Y. Liu, Z. Guo, Y. Zhang, X. Yang, X. Zhang, C. Chen, J. Song, B. Zheng, Y. Yao *et al.*, “Llava-uhd v2: an mllm integrating high-resolution semantic pyramid via hierarchical window transformer,” *arXiv preprint arXiv:2412.13871*, 2024.
- [54] W. Wang, Z. Chen, W. Wang *et al.*, “Enhancing the reasoning ability of multimodal large language models via mixed preference optimization,” *arXiv preprint arXiv:2411.10442*, 2024.
- [55] J. Zhu, W. Wang, Z. Chen *et al.*, “Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models,” *arXiv preprint arXiv:2504.10479*, 2025.
- [56] W. Wang, Z. Gao, L. Gu *et al.*, “Internvl3.5: Advancing open-source multimodal models in versatility, reasoning, and efficiency,” *arXiv preprint arXiv:2508.18265*, 2025.
- [57] L.-C.-T. Xiaomi, “Mimo-vl technical report,” 2025.
- [58] S. Rajbhandari, J. Rasley, O. Ruwase, and Y. He, “Zero: memory optimizations toward training trillion parameter models,” in *Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis*, ser. SC ’20. IEEE Press, 2020.
- [59] T. Dao, “Flashattention-2: Faster attention with better parallelism and work partitioning,” in *The Twelfth International Conference on Learning Representations*, 2024.
- [60] P.-L. Hsu, Y. Dai, V. Kothapalli *et al.*, “Liger-kernel: Efficient triton kernels for LLM training,” in *Championing Open-source DEvelopment in ML Workshop @ ICML25*, 2025.
- [61] W. Kwon, Z. Li, S. Zhuang *et al.*, “Efficient memory management for large language model serving with pagedattention,” in *Proceedings of the 29th Symposium on Operating Systems Principles*, ser. SOSP ’23. New York, NY, USA: Association for Computing Machinery, 2023, p. 611–626.
- [62] Y. Zhan, Z. Xiong, and Y. Yuan, “Rsvg: Exploring data and models for visual grounding on remote sensing data,” *IEEE Transactions on Geoscience and Remote Sensing*, vol. 61, pp. 1–13, 2023.
- [63] X. Li, J. Ding, and M. Elhoseiny, “Vrsbench: A versatile vision-language benchmark dataset for remote sensing image understanding,” *arXiv:2406.12384*, 2024.
- [64] X. Yang, J. Yan, Q. Ming *et al.*, “Rethinking rotated object detection with gaussian wasserstein distance loss,” in *International conference on machine learning*. PMLR, 2021, pp. 11830–11841.
- [65] X. Yang, J. Yan, Z. Feng, and T. He, “R3det: Refined single-stage detector with feature refinement for rotating object,” in *AAAI*, 2021.
- [66] S. Zhang, C. Chi, Y. Yao *et al.*, “Bridging the gap between anchor-based and anchor-free detection via adaptive training sample selection,” in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 9759–9768.
- [67] S. Ren, K. He, R. Girshick, and J. Sun, “Faster r-cnn: Towards real-time object detection with region proposal networks,” *NIPS*, vol. 28, 2015.
- [68] Z. Tian, C. Shen, H. Chen, and T. He, “Fcos: Fully convolutional one-stage object detection,” in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 9627–9636.
- [69] X. Yang and J. Yan, “Arbitrary-oriented object detection with circular smooth label,” in *ECCV*, 2020.
- [70] J. Han, J. Ding, J. Li, and G.-S. Xia, “Align deep features for oriented object detection,” *IEEE transactions on geoscience and remote sensing*, vol. 60, pp. 1–11, 2021.
- [71] J. Chen, D. Zhu, X. Shen *et al.*, “Minigpt-v2: large language model as a unified interface for vision-language multi-task learning,” *arXiv preprint arXiv:2310.09478*, 2023.
- [72] J. A. Irvin, E. R. Liu, J. C. Chen *et al.*, “TEOChat: A large vision-language assistant for temporal earth observation data,” in *The Thirteenth International Conference on Learning Representations*, 2025.
- [73] Z. Li, D. Muhtar, F. Gu *et al.*, “Lhrs-bot-nova: Improved multimodal large language model for remote sensing vision-language interpretation,” *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 227, pp. 539–550, 2025.
- [74] J. Luo, Z. Pang, Y. Zhang *et al.*, “Skysensegt: A fine-grained instruction tuning dataset and model for remote sensing vision-language understanding,” *arXiv preprint arXiv:2406.10100*, 2024.
- [75] D. Muhtar, E. Zhang, Z. Li *et al.*, “Quality-driven curation of remote sensing vision-language data via learned scoring models,” in *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.
- [76] Z. Chen, W. Wang, Y. Cao *et al.*, “Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling,” *arXiv preprint arXiv:2412.05271*, 2024.
- [77] Y. Zhou, X. Yang, G. Zhang *et al.*, “Mmrotate: A rotated object detection benchmark using pytorch,” in *Proceedings of the 30th ACM International Conference on Multimedia*, 2022.
- [78] Y. kelu, X. Nuo, Y. Rong *et al.*, “Falcon: A remote sensing vision-language foundation model,” *arXiv preprint arXiv:2503.11070*, 2025.
- [79] B. Xiao, H. Wu, W. Xu *et al.*, “Florence-2: Advancing a unified representation for a variety of vision tasks,” *arXiv preprint arXiv:2311.06242*, 2023.
- [80] G.-S. Xia, J. Hu, F. Hu *et al.*, “AID: A benchmark data set for performance evaluation of aerial scene classification,” *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 7, pp. 3965–3981, 2017.
- [81] Y. Yang and S. Newsam, “Bag-of-visual-words and spatial extensions for land-use classification,” in *Proceedings of the 18th SIGSPATIAL International Conference on Advances in Geographic Information Systems*, ser. GIS ’10. New York, NY, USA: Association for Computing Machinery, 2010, p. 270–279.
- [82] G. Cheng, J. Han, and X. Lu, “Remote sensing image scene classification: Benchmark and state of the art,” *Proc. of the IEEE*, vol. 105, no. 10, pp. 1865–1883, 2017.
- [83] D. Dai and W. Yang, “Satellite image classification via two-layer sparse coding with biased image representation,” *IEEE Transactions on Geoscience and Remote Sensing*, vol. 8, no. 1, pp. 173–176, 2011.
- [84] B. Zhu, N. Lui, J. Irvin *et al.*, “Meter-ml: a multi-sensor earth observation benchmark for automated methane source mapping,” *arXiv preprint arXiv:2207.11166*, 2022.
- [85] S. Lobry, D. Marcos, J. Murray, and D. Tuia, “Rsvqa: Visual question answering for remote sensing data,” *IEEE Transactions on Geoscience and Remote Sensing*, vol. 58, no. 12, pp. 8555–8566, 2020.