SkyMoE: A Vision-Language Foundation Model for Enhancing Geospatial Interpretation with Mixture of Experts

Jiaqi Liu[1,2], Ronghao Fu[1,2*], Lang Sun[1,2], Haoran Liu[1,2], Xiao Yang[1,2], Weipeng Zhang[1,2], Xu Na[1,2], Zhuoran Duan[1,2], Bo Yang[1,2*]

[1]College of Computer Science and Technology, Jilin University, China
[2]Key Laboratory of Symbolic Computation and Knowledge Engineering of Ministry of Education, Jilin University, China
{liujq21,sunlang24,lhr24,yangx23,zhangwp24,naxu23,duanzr24}@mails.jlu.edu.cn, {furh,ybo}@jlu.edu.cn

Abstract

The emergence of large vision-language models (VLMs) has significantly enhanced the efficiency and flexibility of geospatial interpretation. However, general-purpose VLMs remain suboptimal for remote sensing (RS) tasks. Existing geospatial VLMs typically adopt a unified modeling strategy and struggle to differentiate between task types and interpretation granularities, limiting their ability to balance local detail perception and global contextual understanding. In this paper, we present SkyMoE, a Mixture-of-Experts (MoE) vision-language model tailored for multimodal, multi-task RS interpretation. SkyMoE employs an adaptive router that generates task- and granularity-aware routing instructions, enabling specialized large language model experts to handle diverse sub-tasks. To further promote expert decoupling and granularity sensitivity, we introduce a context-disentangled augmentation strategy that creates contrastive pairs between local and global features, guiding experts toward level-specific representation learning. We also construct MGRS-Bench, a comprehensive benchmark covering multiple RS interpretation tasks and granularity levels, to evaluate generalization in complex scenarios. Extensive experiments on 21 public datasets demonstrate that SkyMoE achieves state-of-the-art performance across tasks, validating its adaptability, scalability, and superior multi-granularity understanding in remote sensing.

Introduction

Recent advances in artificial intelligence have led to the rapid development of Vision-Language Models (VLMs), which demonstrate impressive generalization across a wide range of visual and linguistic tasks (Wang, Wang, and Zhang 2025; Liu et al. 2024b; Durante et al. 2024). Their success is largely attributed to the ability to extract rich, high-level semantic representations from images, enabling capabilities such as open-vocabulary recognition, implicit reasoning, and multimodal alignment. These strengths have sparked growing interest in applying VLMs to Remote Sensing (RS), where tasks such as land cover classification and open-set object detection demand sophisticated scene understanding (Wang et al. 2023; Zhang et al. 2024).

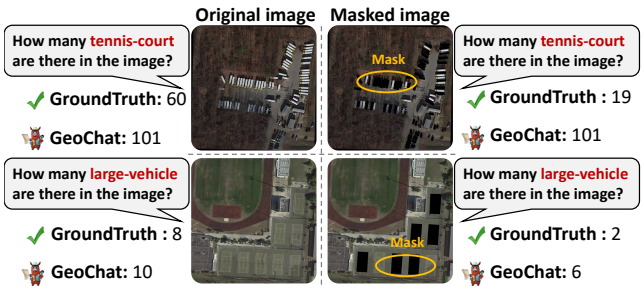

Figure 1: Selective masking of objects in images reveals minimal variation in model-provided counts, indicating a reliance on background context over precise enumeration.

However, RS imagery presents distinct challenges that set it apart from conventional vision tasks. Specifically, it requires a dual capacity to capture global spatial layouts over large geographic extents while simultaneously recognizing fine-grained, localized objects. This coarse-to-fine semantic demand often exceeds the representational limits of standard VLMs, which typically rely on monolithic architectures with shared parameters. Such models struggle to reconcile the competing needs for detailed object-level recognition and holistic scene-level reasoning.

This inherent tension is evident in recent remote sensing vision language models (RS-VLMs). As illustrated in Figure 1, masking-based analyses reveal a strong overreliance on global context. Even when key foreground entities (e.g., vehicles, courts) are occluded, models often produce inflated object counts—indicating a lack of instance-level discrimination and an overdependence on background priors. This behavior highlights a critical shortcoming: the inability to disentangle and properly integrate localized and contextual features, which undermines both model reliability and task-specific performance.

To address this issue, there is increasing interest in modular architectures that support specialization, such as Mixture-of-Experts (MoE) frameworks. MoE architectures offer structural flexibility by distributing computation across expert subnetworks. However, prior RS-based MoE models (Lin et al. 2025) have mostly adopted general-purpose training paradigms without mechanisms to enforce mean-

*Corresponding authors.

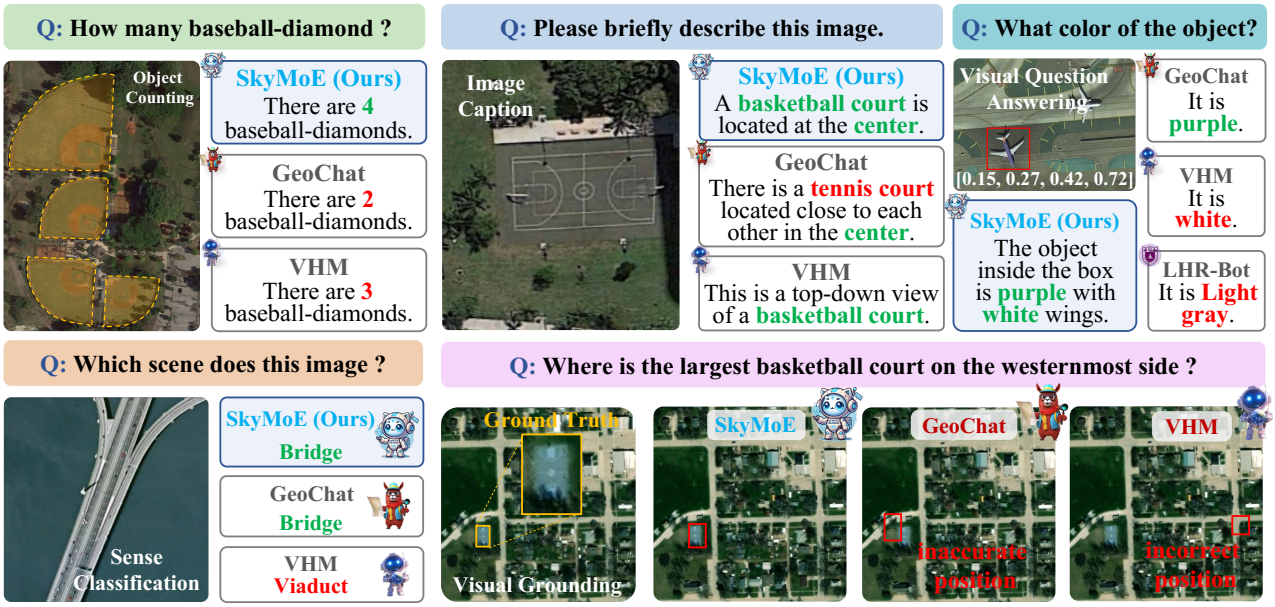

Figure 2: Qualitative results of SkyMoE compared with other remote sensing vision–language models.

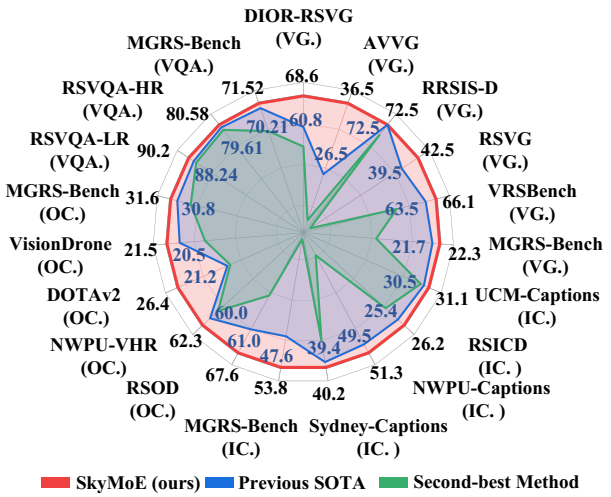

Figure 3: The proposed SkyMoE achieves SOTA performance on 21 benchmarks in 5 remote sensing tasks, outperforming existing RS-VLMs.

ingful expert specialization. As a result, experts often learn overlapping functions, failing to leverage the MoE design for effective feature disentanglement.

In this work, we introduce SkyMoE, a novel VLM framework tailored for multi-scale RS interpretation. SkyMoE integrates two co-designed components: (1) an MoE-based architecture that enables modular specialization, and (2) a contrastive data augmentation strategy that introduces explicit inductive biases to steer expert learning. By systematically modifying local object attributes while preserving global context, we generate fine-grained training samples that compel each expert to focus on either localized or global semantics. This targeted supervision facilitates effective task decomposition and unlocks the MoE architecture's potential for interpretable and robust representation learning. To facilitate comprehensive evaluation, we additionally introduce MGRS-Bench, a dedicated benchmark encompassing a broad range of RS tasks at varying semantic granularities. As shown in Figure 2 and Figure 3, the proposed SkyMoE exhibits competitive performance across multiple tasks. In summary, this work has the following contributions:

- We propose SkyMoE, a novel MoE-based architecture that dynamically routes remote sensing tasks to specialized experts, guided by task-specific interpretation granularity and semantic complexity.
- We design a context-disentangled data augmentation strategy that introduces inductive biases to facilitate expert specialization within the MoE framework.
- We construct MGRS-Bench, a comprehensive benchmark covering diverse tasks, granularity levels, and resolution variations for RS-VLM evaluation.
- Through extensive experiments on 21 diverse benchmarks, we demonstrate that SkyMoE establishes a new state-of-the-art, achieving superior results across a wide spectrum of both fine-grained and scene-level tasks.

Related Work

Vision-Language Models in General Domain

The integration of visual encoders with powerful, open-source Large Language Models (LLMs) has given rise to a new paradigm of Vision-Language Models (VLMs) that have demonstrated remarkable cross-domain capabilities. Models such as LLaVA (Liu et al. 2023b), MiniGPT-4 (Zhu

et al. 2024), and Qwen-VL (Bai et al. 2023) typically employ a unified module to project encoded visual features into the LLM's embedding space. This architectural design, while successful for general-domain tasks (e.g., image captioning, visual dialogue) that prioritize holistic scene understanding, inherently encourages the learning of global contextual features at the expense of fine-grained local details. Consequently, when confronted with specialized domains like remote sensing, which are characterized by dense small objects and demand high-fidelity localization, these general-purpose VLMs often exhibit suboptimal performance. Even advanced models like DeepSeek-VL (Wu et al. 2024), which incorporates an MoE framework, primarily leverage it for scaling model capacity on general tasks, rather than explicitly addressing the fundamental tension between local and global feature representation. This highlights the need for a VLM framework specifically architected and trained for the unique challenges of remote sensing imagery.

Vision-Language Models in Remote Sensing

To address domain-specific challenges, the RS community has developed specialized VLMs. Models like GeoChat (Kuckreja et al. 2024) and EarthGPT (Zhang et al. 2024) have been tailored for geospatial tasks, demonstrating the potential of VLMs in this field. However, these models still largely inherit the architectural tendency of their general-domain counterparts to prioritize global context, thus struggling to achieve a robust balance with local details. Notably, some works have attempted to explicitly address this imbalance. For instance, VHM (Pang et al. 2025) mitigates this issue to an extent by fusing multi-level features from its vision encoder. However, this represents a static, task-agnostic fusion strategy. The combination of features is pre-defined by the architecture and does not adapt to the specific demands of a given input, which may require a dynamic emphasis on either local or global information. This highlights that while the problem of feature balancing is recognized, a mechanism for dynamic, input-aware allocation of model resources remains a critical unfilled gap in RS-VLMs.

Mixture of Experts for Remote Sensing

The promise of MoE as a dynamic, learned framework has recently led to its adoption in the RS domain. For example, RS-MoE (Lin et al. 2025) employs an MoE structure to effectively fuse information from heterogeneous data sources, while RSUniVLM (Liu and Lian 2024) utilizes experts to specialize in different downstream tasks. These pioneering works demonstrate the versatility of the MoE paradigm. However, in these existing frameworks, the MoE mechanism is primarily leveraged for multi-modal fusion or task-level specialization, rather than being explicitly directed at resolving the fundamental, underlying tension between local and global feature representations. Consequently, while they utilize an MoE architecture, their training paradigms are not designed to compel expert differentiation for the specific purpose of feature balancing. As such, the full potential of MoE as a solution to this core challenge remains latent and unrealized, underscoring the novelty and necessity of our proposed approach.

Methodology

The architecture of SkyMoE, as depicted in Figure 4. We begin by describing the architecture in detail. Then, we present the training methodology specifically designed for SkyMoE. Finally, we discuss the training objectives that guide its optimization.

Architecture of SkyMoE

Given an RGB image $\mathbf{v} \in \mathbb{R}^{H\times W\times 3}$, where H and W denote the original resolution, we utilize the pretrained vision backbone of CLIP-ViT(L-14) (Tay et al. 2017), which segments the image into 576 patches at an input resolution of 336×336. To enhance the model's ability to capture intricate details and small objects in RS imagery, we interpolate the positional encoding in the transformer-based CLIP model (Tay et al. 2017) to support input image sizes of 504×504. This enhancement effectively doubles the number of patches to 1296 per image, significantly improving visual grounding in high-resolution RS images.

The vision encoder processes the input image to generate a visual token sequence $\mathcal{Z} = [z_1, z_2, \ldots, z_P] \in \mathbb{R}^{P\times C}$, with $P = \frac{H\times W}{36^2}$ representing the sequence length of visual tokens. A visual projection layer f maps this sequence from $\mathbb{R}^{P\times C}$ to $\mathbb{R}^{P\times D}$, aligning the visual tokens with the hidden size D of the large language model. Concurrently, the text component undergoes a word embedding layer g, resulting in sequence tokens $\mathcal{T} = [t_1, t_2, \ldots, t_N] \in \mathbb{R}^{N\times D}$, where N signifies the sequence length of text tokens.

The visual tokens $\mathcal{V}$ and text tokens $\mathcal{T}$ are concatenated to form a unified sequence $\mathbf{x}_0 = [v_1, v_2, \ldots, v_P, t_1, t_2, \ldots, t_N]$. This combined sequence is then fed into the LLM, which comprises stacked multi-head self-attention (MSA) and feed-forward neural networks (FFN) blocks. Each block incorporates layer normalization (LN) and residual connections, enhancing the model's capacity for complex pattern recognition and feature extraction. The forward pass through the LLM can be mathematically represented as follows:

$$\begin{aligned}\mathbf{x}'_\ell &= \mathrm{MSA}(\mathrm{LN}(\mathbf{x}_{\ell-1})) + \mathbf{x}_{\ell-1}, \quad \ell = 1, \ldots, L,\\ \mathbf{x}_\ell &= \mathrm{MoE}(\mathrm{LN}(\mathbf{x}'_\ell)) + \mathbf{x}'_\ell, \quad \ell = 1, \ldots, L,\end{aligned} \tag{1}$$

where L denotes the number of layers in the LLM. The final output $\mathcal{Y}$ is obtained after the last layer through layer normalization:

$$\mathcal{Y} = \mathrm{LN}(\mathbf{x}_L). \tag{2}$$

The MoE layer, a pivotal component of our model, enhances the LLM's adaptability by incorporating multiple FFNs. Initially, we replicate the FFNs from stage II to form an ensemble of experts $\mathcal{E} = [e_1, e_2, \ldots, e_E]$. A router, implemented as a linear layer, predicts the probability of each token being assigned to each expert. This probability distribution is formulated as:

$$\mathcal{P}(\mathbf{x})_i = \frac{e^{f(\mathbf{x})_i}}{\sum_{j=1}^{E} e^{f(\mathbf{x})_j}}, \tag{3}$$

where the router produces weight logits $f(\mathbf{x}) = \mathbf{W}\cdot\mathbf{x}$, normalized by the softmax function. Here, $\mathbf{W} \in \mathbb{R}^{D\times E}$ represents the lightweight training parameters, and E denotes

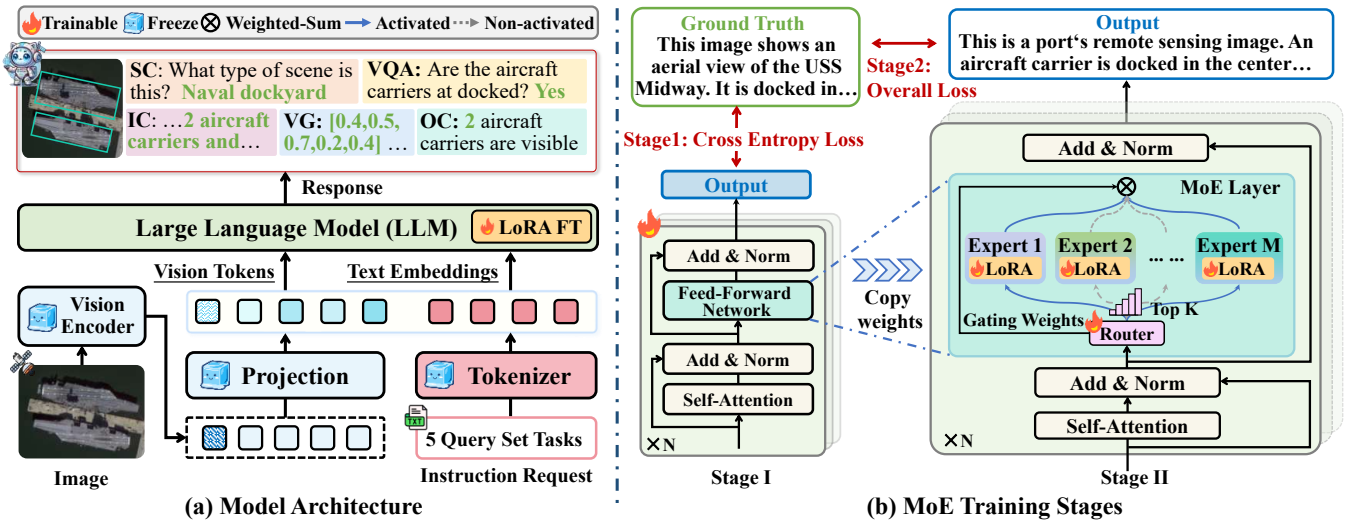

Figure 4: **Training framework and strategy.** The training employs a two-phase approach, Stage I: Initial LLM pretraining establishes multimodal understanding without MoE layers, followed by Stage II: MoE specialization through expert initialization via cloned FFN weights and subsequent fine-tuning.

the number of experts. Each token is processed by the top-k experts with the highest probabilities, and the final output is a weighted sum based on these probabilities:

$$\text{MoE}(\mathbf{x}) = \sum_{i=1}^{k} \mathcal{P}(\mathbf{x})_i \cdot \mathcal{E}(\mathbf{x})_i. \tag{4}$$

Dual-Stage Training

We employ a two-stage strategy for training SkyMoE. Utilizing 6 NVIDIA A800-80G GPUs, we train the model with a batch size of 144 for 5 epochs. The AdamW optimizer (Loshchilov and Hutter 2019) is used with an initial learning rate of 2e-5, combined with a cosine learning rate scheduling strategy. All components of SkyMoE are optimized with the large-scale instruction dataset, the construction of which is introduced in the *Datasets and Evaluation Benchmarks* section, to incorporate RS visual knowledge into the model.

Stage I: To enhance the effectiveness of our model on general visual tasks and optimize training efficiency, we employ a strategy that involves initializing the network with pre-trained weights and fine-tuning specific segments for RS related tasks. Specifically, we utilize a pre-trained CLIP-ViT(L-14) encoder (Tay et al. 2017), trained on large amounts of textual and visual data, a pre-trained MLP adaptor (Liu et al. 2024a), trained on a 558K subset of the LAION-CC-SBU dataset (Schuhmann et al. 2022) with BLIP (Li et al. 2022) captions, and Vicuna-v1.5 (Shen et al. 2023) to initialize our model. To adapt our model to RS images, we subsequently apply LoRA (Hu et al. 2022) fine-tune on the LLM, targeting the W_q and W_v matrices with a designated rank r set to 64, while keeping the MLP adaptor and the CLIP encoder (Tay et al. 2017) frozen during training. The model undergoes training consistently at an image resolution of 504 $\times$ 504 throughout the whole process. This stage not only lays the foundation for the model's ability to handle general visual tasks but also enhances its capabilities and controllability by tuning it with multi-modal instruction data (Kuckreja et al. 2024). We use more complex instructions, including tasks such as image logical reasoning and text recognition, which require the model to have a stronger multi-modal understanding. Typically, for dense models, the VLM training is considered complete at this stage. However, we encounter challenges in simultaneously transforming the LLM into an VLM and sparsifying the VLM. To address this, SkyMoE utilizes the weights from this stage as initialization for the next stage to alleviate the learning difficulty of the sparse model.

Stage II: In the second stage, we transform the model into a sparse VLM by integrating a MoE architecture. Expert modules are initialized by replicating the FFN multiple times, and MoE layers are interleaved with standard MLP layers. Specifically, every second layer is replaced with an MoE layer. When image tokens and text tokens are fed into the MoE layers, the router calculates the matching weights between each token and the experts. Each token is then processed by the top-2 experts, and the outputs are aggregated by weighted summation based on the router's weights. Non-selected experts remain inactive, enabling dynamic sparsity and efficient computation. Expert parallel size is set to 1 to support distributed expert execution. This architecture enables SkyMoE to adaptively select computation paths, achieving both high-level reasoning and fine-grained perceptual understanding across diverse RS tasks.

Training Objectives

The total loss function $\mathcal{L}_{\text{total}}$ is composed of the auto-regressive loss $\mathcal{L}_{\text{regressive}}$ and the auxiliary loss $\mathcal{L}_{\text{aux}}$, with the auxiliary loss being scaled by a balancing coefficient α:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{regressive}} + \alpha \cdot \mathcal{L}_{\text{aux}}. \tag{5}$$

Auto-Regressive Loss. The output of the LLM is optimized via a generative loss in an auto-regressive fashion. Given an image and text, SkyMoE produces the output sequence $\mathcal{Y} = [y_1, y_2, \ldots, y_K] \in \mathbb{R}^{K \times D}$ by sequentially generating each element, where $K = P + N$ denotes the length of the output sequence. The loss function is defined as:

$$\mathcal{L}_{\text{regressive}} = -\sum_{i=1}^{N} \log\, p_\theta \left(\mathcal{Y}^{[P+i]} \mid \mathcal{V}, \mathcal{T}^{[:i-1]} \right), \tag{6}$$

where θ represents the trainable parameters, and the loss is computed solely for the generated text.

Auxiliary Loss. Given the presence of multiple experts, it is essential to enforce load balancing constraints on the MoE layer. We integrate a differentiable load balancing loss into each MoE layer to promote balanced token handling among experts:

$$\mathcal{L}_{\text{aux}} = E \cdot \sum_{i=1}^{E} \mathcal{F}_i \cdot \mathcal{G}_i, \tag{7}$$

where $\mathcal{F}$ denotes the fraction of tokens processed by each expert $\mathcal{E}_i$, and $\mathcal{G}$ represents the average routing probability for $\mathcal{E}_i$, which are given by:

$$\mathcal{F} = \frac{1}{K} \sum_{i=1}^{E} \mathbb{1}\{\text{argmax}\, \mathcal{P}(\mathbf{x}) = i\}, \mathcal{G} = \frac{1}{K} \sum_{i=1}^{K} \mathcal{P}(\mathbf{x})_i. \tag{8}$$

Datasets and Evaluation Benchmarks

To support the training and evaluation of SkyMoE, we construct a comprehensive dataset pipeline encompassing both foundational data aggregation and targeted augmentation. In Stage I, we aggregate diverse RS datasets spanning multiple tasks to form a unified instruction-following training corpus. In Stage II, we apply task-specific augmentation strategies (Figure 5) to enhance data diversity and optimize the MoE fine-tuning process. Together, these stages ensure that SkyMoE is exposed to a wide range of task types, object densities, and attribute variations—ultimately boosting its generalization ability in real-world RS scenarios. In addition, we introduce MGRS-Bench, a purpose-built benchmark that enables multi-granularity and cross-resolution evaluation, thereby filling critical gaps in existing test sets.

Instruction Dataset Construction

To enable SkyMoE to handle diverse RS tasks in a unified manner, we construct a multi-task instruction dataset by curating and integrating multiple high-quality, publicly available sources. The dataset spans five vision-language tasks. Our final training set includes over 251k instruction samples, with an additional 11k instances for testing. The training data is derived from established benchmarks such as DOTA (Xia et al. 2018), DIOR (Li et al. 2020), FAIR1M (Sun et al. 2022a) , DOTA-v2 (Xia et al. 2018). To ensure robust generalization, the test set is sourced independently from datasets like DIOR-RSVG (Zhan, Xiong, and Yuan 2023), RSOD (Long et al. 2017), and NWPU-RESISC45-test (Cheng, Han, and Lu 2017). A full list of source datasets is available in the supplementary material.

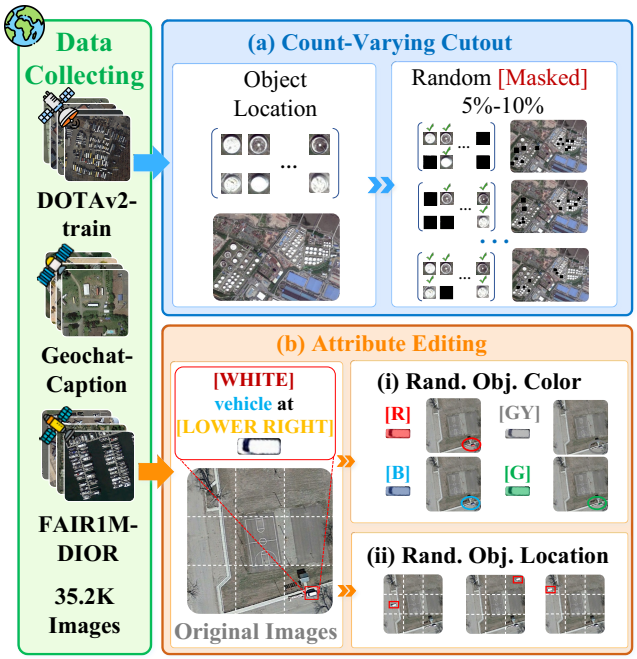

Figure 5: Context-Disentangled Data Augmentation.

To unify training across tasks, we reformat each sample into an instruction-following format. For this, we design task-specific templates that convert original annotations into natural language instructions and standardized answers. For example, scene classification is expressed as *"Which scene does this image belong to?"*, expecting a category name as output. Visual grounding uses the prompt *"[refer] Where is referring expression ?"*, requiring bounding box coordinates. Similar templates are used for counting, captioning, and VQA. This unified structure allows the model to learn diverse tasks under a single conversational paradigm.

Context-Disentangled Data Augmentation

To enhance the diversity and granularity sensitivity of the training data, we design augmentation strategies that explicitly vary object count, position, and appearance—encouraging SkyMoE to specialize across experts and better capture multi-scale semantics.

Count-Varying Cutout. To simulate diverse object densities for counting tasks, we randomly mask a portion of objects in each image. Given an original count N_c, we sample a masking ratio $r \sim \mathcal{U}(0.15, 0.30)$ and remove $m = \lceil r \cdot N_c \rceil$ objects by zero-filling their regions. Updated annotations reflect the new count $N_{\text{new}} = N_c - m$. Each image yields four variants with at least $\lceil 0.1 N_c \rceil$ difference in count, enabling fine-grained supervision for density-sensitive predictions.

Attribute Editing. To strengthen vision-language alignment, we manipulate referring expressions by editing spatial and color attributes, then apply corresponding visual transformations. For spatial editing, positional phrases (e.g., "bottom") are replaced (e.g., "top right") and the target object is relocated within a 3×3 grid cell, ensuring visual real-

Method	UCM-Captions			RSICD			RSITMD			NWPU-Captions			Sydney-Captions			MGRS-Bench		
	B-4	MT	R-L	B-4	MT	R-L	B-4	MT	R-L	B-4	MT	R-L	B-4	MT	R-L	B-4	MT	R-L
Close-source Commercial Vision-Language Models																		
GPT-4V [OpenAI'24]	28.4	25.6	17.8	16.8	16.7	15.9	27.3	21.1	14.0	39.6	25.7	20.9	28.5	24.5	17.5	22.1	32.6	22.3
Open-source Vision-Language Models																		
MiniGPT-v2 [Meta'24]	18.1	21.7	11.4	11.9	15.8	11.2	19.0	19.7	10.0	28.6	23.3	13.0	19.8	19.5	10.6	19.4	15.9	16.9
LLaVA-1.5 [CVPR'24]	24.6	26.7	17.1	21.3	16.4	17.9	35.0	23.1	16.7	47.6	26.5	21.9	39.1	25.3	21.1	36.6	28.2	36.0
Qwen2.5-VL [BABA'24]	14.7	24.5	11.5	9.5	17.2	11.2	16.6	21.8	10.5	24.1	26.4	14.4	16.1	22.5	11.3	30.7	47.6	**37.8**
DeepSeek-VL [DeepSeek'24]	23.1	30.5	16.6	13.9	20.6	14.6	26.9	25.0	14.6	39.1	30.5	20.4	38.2	26.1	20.7	31.6	28.0	33.3
Open-source Remote Sensing Vision-Language Models																		
GeoChat [CVPR'24]	21.0	20.8	14.1	15.8	14.1	13.4	27.5	20.3	13.8	37.8	25.2	16.3	24.6	16.5	10.5	24.4	13.9	22.7
VHM [AAAI'25]	42.4	26.9	**24.9**	19.5	22.2	18.0	36.4	17.9	15.7	47.9	18.9	17.8	37.0	21.5	17.4	29.1	21.9	28.3
SkySenseGPT [arXiv'24]	15.7	23.3	10.9	12.2	17.0	12.8	20.6	21.7	11.8	28.2	24.7	14.4	18.4	21.3	11.9	24.9	17.2	25.2
LHRS-Bot [ECCV'24]	23.5	30.1	17.1	21.2	23.8	19.4	**38.2**	27.9	**19.8**	58.4	49.5	35.8	40.2	36.2	25.1	25.6	20.8	25.2
RSUniVLM [arXiv'24]	18.3	19.5	11.5	10.6	13.5	11.4	19.8	20.8	11.3	26.7	21.0	13.9	20.2	19.8	10.6	15.6	9.6	16.5
Falcon [arXiv'25]	14.4	6.0	9.9	1.3	0.9	3.4	9.2	5.3	8.9	3.0	1.0	0.8	3.6	1.9	2.7	12.9	4.2	7.6
EarthDial [CVPR'25]	14.6	18.1	10.0	27.8	25.4	**24.1**	23.0	19.2	13.3	35.9	28.8	18.4	45.0	39.4	**25.9**	25.8	14.8	20.4
SkyMoE (ours)	**43.0**	**31.1**	19.7	**33.4**	**26.2**	19.6	37.7	**28.3**	18.6	**60.1**	**51.3**	**38.8**	**46.1**	**40.2**	25.2	**38.7**	**53.8**	37.7

Table 1: Comparison of SkyMoE with existing generic and RS VLMs on Image Captioning task across multiple benchmarks. B-4, MT, and R-L denote BLUE-4, METEOR, and ROUGE-L scores, respectively.

ism via Poisson blending and minimal overlap (IoU < 0.1). For color editing, textual color terms are swapped (e.g., "green" $\rightarrow$ "red"), and matched via palette transfer while preserving material texture and reflectance.

These granularity-aware augmentations diversify scene configurations while maintaining annotation consistency, allowing SkyMoE's experts to specialize in distinct visual patterns and improving overall generalization.

Multi-Granularity RS Benchmark

To address the limitations of existing RS-VLM benchmarks, such as limited task diversity, insufficient granularity control, and restricted resolution variation, we introduce MGRS-Bench, a dedicated evaluation suite for RS-VLMs.

MGRS-Bench is constructed through a multi-stage annotation pipeline consisting of four components: (1) attribute extraction to identify relevant visual features and contextual cues, (2) prompt engineering to generate diverse instruction formats, (3) GPT-4 inference to produce high-quality candidate responses, and (4) human verification to ensure accuracy, disambiguation, and label consistency. The final dataset includes 10,415 RS images and 18,433 annotated instances, with 2,083 images assigned to each of the five tasks.

Experiments

Evaluation Results

To thoroughly assess the performance of various VLMs in RS tasks, we employ a comprehensive evaluation scheme that includes Image Captioning (IC), Visual Question Answering (VQA), Visual Grounding (VG), Object Counting (OC), and Scene Classification (SC). This evaluation scheme is designed to provide a holistic view of each model's capabilities across a spectrum of tasks, from high-level reasoning to detailed perceptual tasks.

Image Captioning. SkyMoE achieves competitive performance in IC, demonstrating robust generalization across six public benchmarks (Lu et al. 2017; Yuan et al. 2021; Chen et al. 2022; Qu et al. 2016). As shown in Table 1, it obtains the highest BLEU-4 scores on UCM-Captions (43.0), RSICD (33.4), and NWPU-Captions (60.1), and leads in METEOR on six datasets, including a strong 53.8 on MGRS-Bench. These results highlight SkyMoE's superior ability to generate accurate, fluent, and context-aware descriptions. Compared to existing RS-VLMs, most of which can recognize object pairs or isolated features, SkyMoE demonstrates a clear advantage in modeling complex spatial and semantic relations. For example, while SkySenseGPT performs competitively in grounding tasks, its performance on RSICD and MGRS-Bench captioning drops significantly (e.g., BLEU-4: 12.2 and 24.9), suggesting a limited capacity to structure local features into coherent narratives. In contrast, SkyMoE's consistent gains across BLEU, ME-

Method	RSVQA-LR			RSVQA-HR		MGRS-Bench	
	LR-R	LR-P	LR-C	HR-A	HR-C	MG-P	MG-D
Close-source Commercial Vision-Language Models							
GPT-4V [OpenAI'24]	95.90	49.95	51.95	0	68.93	92.41	42.37
Open-source Vision-Language Models							
MiniGPT-v2	54.72	48.21	60.38	0	59.22	58.82	40.43
LLaVA-1.5	75.47	57.14	70.59	0	66.02	88.24	63.83
Qwen2.5-VL	94.34	62.50	43.14	0	53.40	70.59	70.21
DeepSeek-VL	62.26	48.21	64.71	0	67.96	85.29	34.04
Open-source Remote Sensing Vision-Language Models							
GeoChat	96.23	89.29	88.24	23.30	75.73	94.12	63.83
VHM	90.57	80.36	86.27	24.27	79.61	88.24	42.55
SkySenseGPT	92.45	85.71	84.31	0	78.64	**95.06**	63.8
LHRS-Bot	84.91	76.79	49.02	0	37.86	94.32	53.19
RSUniVLM	35.85	73.21	65.69	0	76.70	85.14	27.66
Falcon	58.75	45.95	59.94	0	42.96	68.7	42.12
EarthDial	66.04	64.29	87.25	**34.95**	71.84	76.47	46.81
SkyMoE (ours)	**98.11**	**91.07**	**90.20**	33.98	**80.58**	94.38	**71.52**

R/P/C/A/D: Rural / Presence / Comparison / Area / Direction.

Table 2: Comparison of SkyMoE with existing generic and RS VLMs on VQA task.

Method	DIOR-RSVG		AVVG		RRSIS-D		RSVG		VRSBench		MGRS	
	@0.5	@0.7	@0.5	@0.7	@0.5	@0.7	@0.5	@0.7	@0.5	@0.7	@0.5	@0.7
Close-source Commercial Vision-Language Models												
GPT-4V [OpenAI'24]	26.1	3.6	33.2	8.7	28.0	5.0	36.5	18.3	14.4	2.3	18.3	7.6
Open-source Vision-Language Models												
MiniGPT-v2	12.2	5.3	1.5	1.0	14.0	6.5	0	0	12.4	5.7	3.9	0.8
LLaVA-1.5	11.4	1.6	0.5	0	12.0	2.0	10.5	2.5	15.4	5.6	19.4	5.4
Qwen2.5-VL	36.3	19.2	0.5	0	0.5	0	1.0	0	45.2	25.9	1.6	0
Open-source Remote Sensing Vision-Language Models												
GeoChat	31.4	14.7	7.5	0.5	10	1.5	5.5	1	56.3	32.6	17.1	10.1
VHM	55.9	42.0	0	0	64.0	45.0	2.5	0.5	33.9	14.3	21.7	10.9
SkySenseGPT	60.8	35.5	26.5	3.5	69.0	41.5	39.5	21.5	63.5	34.2	3.1	0.7
GeoGround	37.6	22.0	20.0	9.0	59.5	39.5	22.5	11.0	41.4	21.6	7.0	2.3
EarthDial	46.1	34.3	3.5	2.0	**72.5**	55.5	42.0	24.0	14.4	9.7	17.1	9.3
SkyMoE (Ours)	**68.6**	**48.6**	**36.5**	**18.0**	**72.5**	**56.1**	**42.5**	**26.0**	**66.1**	**44.4**	**22.3**	**13.4**

Table 3: Comparison of SkyMoE with generic and RS VLMs on Visual Grounding task using mAP@0.5 and @0.7.

Method	RSOD	HRRSD	VHR	DOTA	VisDrone	MGRS
Close-source Commercial Vision-Language Models						
GPT-4V [OpenAI'24]	40.0	58.5	58.0	19.7	19.0	28.4
Open-source Vision-Language Models						
MiniGPT-v2	20.0	26.5	23.0	7.7	8.5	4.6
LLaVA-1.5	42.5	52.4	40.5	11.8	10.5	21.5
Qwen2.5-VL	40.5	56.7	57.5	17.4	14.0	27.7
DeepSeek-VL	61.0	**61.5**	60.0	18.3	20.5	29.2
Open-source Remote Sensing Vision-Language Models						
GeoChat	19.5	57.6	42.5	16.9	14.5	21.5
VHM	16.0	46.7	48.5	18.0	5.5	30.8
SkySenseGPT	51.5	58.7	49.5	21.2	18.5	27.7
LHRS-Bot	2.0	2.3	3.0	10.5	1.5	12.3
RSUniVLM	49.5	54.2	36.0	19.0	18.5	27.5
Falcon	47.0	42.6	52.5	18.1	3.5	26.4
EarthDial	41.0	61.5	52.5	20.9	18.0	29.2
SkyMoE (ours)	**67.6**	57.8	**62.3**	**26.4**	**21.5**	**31.6**

Table 4: Comparison of SkyMoE for Object Counting tasks across various datasets.

Method	VG (mAP@0.5)	OC (Acc)	IC (BLEU-4)	VQA (Acc)	SC (Acc)
RSUniVLM	17.1	27.7	15.6	56.4	7.6
MoE-LLaVA	18.9	28.2	20.7	61.0	11.6
SkyMoE (ours)	**22.3**	**31.6**	**38.7**	**83.0**	**68.7**

Table 5: Comparison with other MoE-based VLMs.

TEOR, and ROUGE-L indicate its holistic reasoning capability, empowered by MoE-based expert specialization and fine-grained data augmentation.

Visual Question Answering. As shown in Table 2, SkyMoE consistently outperforms all baselines across diverse VQA scenarios (Lobry et al. 2020). It excels on both low-resolution tasks that require global contextual understanding (e.g., 98.11% on rural classification) and high-resolution tasks that demand fine-grained spatial precision (e.g., 33.98% on area estimation). Notably, SkyMoE maintains strong performance across resolution levels, as seen in the comparison tasks (90.20% on LR-C vs. 80.58% on HR-C), indicating its robust ability to balance local and global feature representations. The resolution gap in RS-VQA fundamentally tests a model's capacity to generalize across different perceptual scales: low-resolution questions rely on coarse semantic cues, whereas high-resolution ones require precise detail recognition. SkyMoE effectively bridges this gap through the synergy between its MoE architecture and context-disentangled augmentation strategy. This dynamic adaptation mechanism allows the model to tailor its representation to task-specific granularity, setting a new benchmark for resolution-variant RS VQA.

Visual Grounding. SkyMoE delivers consistent and competitive performance across all VG benchmarks (Zhan, Xiong, and Yuan 2023; Zhou et al. 2025; Liu et al. 2024c; Sun et al. 2022b; Li, Ding, and Elhoseiny 2024). As shown in Table 3, it achieves a strong mAP@0.5 of 68.6% on DIOR-RSVG, with only moderate degradation at mAP@0.7 (48.6%), indicating its robust ability to balance global scene comprehension with precise object localization. In contrast, generic VLMs like MiniGPT-v2 and LLaVA-1.5 struggle to surpass even 12% at mAP@0.5 and collapse almost entirely at stricter thresholds (e.g., MiniGPT-v2: 12.2% $\rightarrow$ 5.3%), reflecting over-reliance on global semantics while lacking spatial precision. RS-VLMs (e.g., GeoChat, VHM, SkySenseGPT) exhibit stronger low-threshold performance (e.g., SkySenseGPT: 60.8% mAP@0.5 on DIOR-RSVG), but their performance degrades significantly at mAP@0.7 (e.g., SkySenseGPT: 35.5%), revealing a deficiency in handling fine-grained grounding. On AVVG, which features high spatial ambiguity, SkyMoE further distinguishes itself by maintaining 36.5% (mAP@0.5) and 18.0% (mAP@0.7), while most baselines fall below 5%. These results underscore SkyMoE's robustness in integrating global and local cues, a key capability for real-world RS applications.

Object Counting. As shown in Table 4, SkyMoE demonstrates strong density-invariant performance, excelling in both sparse (e.g., 67.6% on RSOD) and ultra-dense (e.g., 26.4% on DOTAv2) aerial scenes (Long et al. 2017; Zhang

Dataset	Open-source Vision-Language Models				Open-source Remote Sensing Vision-Language Models							
	MiniGPT-v2	LLaVA-1.5	Qwen2.5-VL	DeepSeek-VL	GeoChat	VHM	SkySenseGPT	LHRS-Bot	RSUniVLM	EarthDial	Falcon	**SkyMoE**
RESISC45	21.78	42.22	69.78	46.00	84.67	91.33	83.33	42.44	2.22	76.67	62.22	**91.77**
AID	18.83	47.33	63.83	34.67	71.50	**79.00**	75.50	39.00	3.33	62.50	20.83	78.21
WHU-RS19	26.84	62.11	79.47	80.00	89.47	91.84	**93.16**	56.58	5.26	73.42	44.21	92.37
MGRS-Bench	35.10	59.35	**76.44**	44.80	56.35	66.28	55.89	51.04	7.62	42.96	38.89	68.73

Table 6: Comparison of SkyMoE for Scene Classification tasks across multiple benchmarks.

Metric	GeoChat	VHM	SkySense	LHRS-Bot	Ours
Param Size	7.06B	6.77B	7.05B	6.74B	9.36B
TFLOPs	694.31	665.30	694.31	662.41	712.56
Latency	26.92	21.92	24.59	**18.97**	20.83

Table 7: Results on training and inference efficiency.

#E	CDA		VG	IC	OC	VQA	SC
	CVC	AE	(mAP@0.5)	UCM-Cap	(Acc)	(Acc)	(Acc)
Varying number of experts (with both modules enabled)							
1	✓	✓	45.40	45.4	44.50	74.67	69.67
4	✓	✓	63.79	67.4	58.49	83.64	80.20
6	✓	✓	67.64	73.2	66.65	91.51	89.26
8	✓	✓	**68.60**	**78.8**	**68.51**	**93.13**	**91.77**
Ablating key components (fixed #Expert = 8)							
8	✗	✗	60.20	72.0	56.37	82.54	91.29
8	✓	✗	60.34	72.1	65.84	89.88	**91.86**
8	✗	✓	68.10	77.6	61.23	86.57	90.11
8	✓	✓	**68.60**	**78.8**	**68.51**	**93.13**	91.77

CDA: Context-Disentagled Augumentation, CVC: Count-Varying Cutout, AE: Attribute Editing

Table 8: Performance with different architectures.

et al. 2019; Cheng et al. 2014; Xia et al. 2018; Zhu et al. 2021), showcasing a remarkable ability to dynamically balance fine-grained local details and global contextual cues. In sparse environments, it precisely detects individual instances, while in dense scenes, it preserves instance awareness without losing structural coherence. This robustness stems from the synergy of our MoE architecture and the proposed count-varying cutout augmentation, which explicitly trains the model to adapt to density shifts. SkyMoE also achieves a new state-of-the-art on NWPU-VHR (62.3%), further establishing its superiority in the challenging task of aerial object counting.

Scene Classification. SkyMoE achieves highly competitive results on scene classification benchmarks (Cheng, Han, and Lu 2017; Xia et al. 2017, 2010) (e.g., 78.21% on AID, 92.37% on WHU-RS19), all within 2% of the top-performing models. This small gap reflects a deliberate design trade-off: unlike models that solely optimize for global semantics, SkyMoE's MoE framework allocates capacity to preserve fine-grained local features. While slightly limiting performance on purely global tasks, it enables SkyMoE to outperform all baselines on more demanding tasks like visual grounding and object counting, resulting in a more robust and versatile model for diverse geospatial challenges.

Comparison with MoE-based VLMs. As shown in Table 5, SkyMoE achieves the best performance across all five RS vision-language tasks on MGRS-Bench, significantly outperforming previous MoE-based models such as RSUniVLM and MoE-LLaVA. SkyMoE achieves over +18 BLEU-4 improvement in IC. The substantial gains in high-level reasoning tasks like VQA (+22.0 Acc) and SC (+57.1 Acc) further demonstrate the model's strong capability in both fine-grained perception and global semantic understanding.

Training and Inference Efficiency. We compare the training and inference costs of SkyMoE with four recent RS-VLMs in Table 7. While SkyMoE has a larger number of trainable parameters, its overall training complexity remains comparable to other models. It also achieves competitive per-token latency, outperforming larger models such as GeoChat and SkySenseGPT. This efficiency is attributed to the top-k expert activation mechanism, which preserves model capacity while reducing inference overhead. Compared with LHRS-Bot's fully shared architecture, SkyMoE offers a better trade-off between scalability and efficiency.

Ablation study. Our ablation study dissects SkyMoE to validate the contribution of its components. First, we demonstrate the efficacy of the MoE architecture itself, as performance scales positively with the number of experts (e.g., VG improves from 45.4% to 68.6% as experts increase from 1 to 8). Second, our purpose-built data augmentation strategy provides a significant performance uplift, proving particularly effective for tasks sensitive to local details such as object counting (+9.47%). Critically, the full framework significantly outperforms any partial configuration. For example, the complete model achieves a VQA score of 93.13%, a +6.56 gain over using the MoE architecture with standard training. This non-additive performance gain proves that our SOTA results are not achieved through isolated improvements, but through a co-designed framework where our targeted data augmentation actively unlocks and directs the latent potential of the MoE architecture.

Conclusion

In this work, we proposed SkyMoE, a novel framework that combines a MoE architecture with a tailored data augmentation strategy to promote expert specialization. Unlike prior methods, SkyMoE dynamically allocates experts based on feature granularity, enabling a robust balance between local and global representations. To support systematic evaluation, we constructed MGRS-Bench, a comprehensive benchmark covering diverse RS tasks and granularity levels. Extensive experiments on 21 datasets demonstrate SOTA performance, with ablations and visualizations confirming that this improvement stems from our model's learned ability to adaptively route information.

Acknowledgments

This work was supported by the National Key Research and Development Program of China (Grant No. 2021ZD0112500), National Natural Science Foundation of China (Grant No. U22A2098, 62172185, 62206105 and 62202200), Major Science and Technology Project of Jilin Province (Grant No. 20240212001GX), and Major Science and Technology Project of Changchun City (Grant No. 2024WX05).

References

Bai, J.; Bai, S.; Yang, S.; Wang, S.; Tan, S.; Wang, P.; Lin, J.; Zhou, C.; and Zhou, J. 2023. Qwen-VL: A Versatile Vision-Language Model for Understanding, Localization, Text Reading, and Beyond. arXiv:2308.12966.

Bai, S.; Chen, K.; Liu, X.; Wang, J.; Ge, W.; Song, S.; Dang, K.; Wang, P.; Wang, S.; Tang, J.; Zhong, H.; Zhu, Y.; Yang, M.; Li, Z.; Wan, J.; Wang, P.; Ding, W.; Fu, Z.; Xu, Y.; Ye, J.; Zhang, X.; Xie, T.; Cheng, Z.; Zhang, H.; Yang, Z.; Xu, H.; and Lin, J. 2025. Qwen2.5-VL Technical Report. arXiv:2502.13923.

Banerjee, S.; and Lavie, A. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, 65–72.

Chen, J.; Zhu, D.; Shen, X.; Li, X.; Liu, Z.; Zhang, P.; Krishnamoorthi, R.; Chandra, V.; Xiong, Y.; and Elhoseiny, M. 2023. MiniGPT-v2: large language model as a unified interface for vision-language multi-task learning. arXiv:2310.09478.

Chen, L.; Zhang, Y.; Tian, B.; Ai, Y.; Cao, D.; and Wang, F.-Y. 2022. Parallel driving OS: A ubiquitous operating system for autonomous driving in CPSS. *IEEE Transactions on Intelligent Vehicles*, 7(4): 886–895.

Cheng, G.; Han, J.; and Lu, X. 2017. Remote sensing image scene classification: Benchmark and state of the art. *Proceedings of the IEEE*, 105(10): 1865–1883.

Cheng, G.; Han, J.; Zhou, P.; and Guo, L. 2014. Multi-class geospatial object detection and geographic image classification based on collection of part detectors. *ISPRS Journal of Photogrammetry and Remote Sensing*, 98: 119–132.

Durante, Z.; Huang, Q.; Wake, N.; Gong, R.; Park, J. S.; Sarkar, B.; Taori, R.; Noda, Y.; Terzopoulos, D.; Choi, Y.; Ikeuchi, K.; Vo, H.; Fei-Fei, L.; and Gao, J. 2024. Agent AI: Surveying the Horizons of Multimodal Interaction. arXiv:2401.03568.

Hu, E. J.; Shen, Y.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; and Chen, W. 2022. LoRA: Low-Rank Adaptation of Large Language Models. In *International Conference on Learning Representations*.

kelu, Y.; Nuo, X.; Rong, Y.; Yingying, X.; Zhuoyan, G.; Titinunt, K.; yi, R.; Pu, Z.; Jin, W.; Ning, W.; and Chao, L. 2025. Falcon: A Remote Sensing Vision-Language Foundation Model. *arXiv preprint arXiv:2503.11070*.

Kuckreja, K.; Danish, M. S.; Naseer, M.; Das, A.; Khan, S.; and Khan, F. S. 2024. Geochat: Grounded large vision-language model for remote sensing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 27831–27840.

Li, J.; Li, D.; Xiong, C.; and Hoi, S. 2022. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*, 12888–12900. PMLR.

Li, K.; Wan, G.; Cheng, G.; Meng, L.; and Han, J. 2020. Object detection in optical remote sensing images: A survey and a new benchmark. *ISPRS journal of photogrammetry and remote sensing*, 159: 296–307.

Li, X.; Ding, J.; and Elhoseiny, M. 2024. VRSBench: A Versatile Vision-Language Benchmark Dataset for Remote Sensing Image Understanding. In *Advances in Neural Information Processing Systems*, volume 37, 3229–3242. Curran Associates, Inc.

Lin, C.-Y. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, 74–81.

Lin, H.; Hong, D.; Ge, S.; Luo, C.; Jiang, K.; Jin, H.; and Wen, C. 2025. Rs-moe: A vision-language model with mixture of experts for remote sensing image captioning and visual question answering. *IEEE Transactions on Geoscience and Remote Sensing*.

Liu, H.; Li, C.; Li, Y.; and Lee, Y. J. 2023a. Llava-1.5: Improved baselines with visual instruction tuning.

Liu, H.; Li, C.; Li, Y.; and Lee, Y. J. 2024a. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 26296–26306.

Liu, H.; Li, C.; Wu, Q.; and Lee, Y. J. 2023b. Visual instruction tuning. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23. Red Hook, NY, USA: Curran Associates Inc.

Liu, J.; Qi, X.; Hang, P.; and Sun, J. 2024b. Enhancing social decision-making of autonomous vehicles: A mixed-strategy game approach with interaction orientation identification. *IEEE Transactions on Vehicular Technology*.

Liu, S.; Ma, Y.; Zhang, X.; Wang, H.; Ji, J.; Sun, X.; and Ji, R. 2024c. Rotated multi-scale interaction network for referring remote sensing image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 26658–26668.

Liu, X.; and Lian, Z. 2024. Rsunivlm: A unified vision language model for remote sensing via granularity-oriented mixture of experts. *arXiv preprint arXiv:2412.05679*.

Lobry, S.; Marcos, D.; Murray, J.; and Tuia, D. 2020. RSVQA: Visual question answering for remote sensing data. *IEEE Transactions on Geoscience and Remote Sensing*, 58(12): 8555–8566.

Long, Y.; Gong, Y.; Xiao, Z.; and Liu, Q. 2017. Accurate object localization in remote sensing images based on convolutional neural networks. *IEEE Transactions on Geoscience and Remote Sensing*, 55(5): 2486–2498.

Loshchilov, I.; and Hutter, F. 2019. Decoupled Weight Decay Regularization. arXiv:1711.05101.

Lu, X.; Wang, B.; Zheng, X.; and Li, X. 2017. Exploring models and data for remote sensing image caption generation. *IEEE Transactions on Geoscience and Remote Sensing*, 56(4): 2183–2195.

Luo, J.; Pang, Z.; Zhang, Y.; Wang, T.; Wang, L.; Dang, B.; Lao, J.; Wang, J.; Chen, J.; Tan, Y.; and Li, Y. 2024. SkySenseGPT: A Fine-Grained Instruction Tuning Dataset and Model for Remote Sensing Vision-Language Understanding. arXiv:2406.10100.

Muhtar, D.; Li, Z.; Gu, F.; Zhang, X.; and Xiao, P. 2024. Lhrs-bot: Empowering remote sensing with vgi-enhanced large multimodal language model. In *European Conference on Computer Vision*, 440–457. Springer.

OpenAI. 2023. GPT-4V system card. https://openai.com/index/gpt-4v-system-card.

Pang, C.; Weng, X.; Wu, J.; Li, J.; Liu, Y.; Sun, J.; Li, W.; Wang, S.; Feng, L.; Xia, G.-S.; et al. 2025. Vhm: Versatile and honest vision language model for remote sensing image analysis. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 6381–6388.

Papineni, K.; Roukos, S.; Ward, T.; and Zhu, W.-J. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, 311–318.

Qu, B.; Li, X.; Tao, D.; and Lu, X. 2016. Deep semantic understanding of high resolution remote sensing image. In *2016 International conference on computer, information and telecommunication systems (Cits)*, 1–5. IEEE.

Schuhmann, C.; Beaumont, R.; Vencu, R.; Gordon, C.; Wightman, R.; Cherti, M.; Coombes, T.; Katta, A.; Mullis, C.; Wortsman, M.; et al. 2022. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in neural information processing systems*, 35: 25278–25294.

Shen, Y.; Song, K.; Tan, X.; Li, D.; Lu, W.; and Zhuang, Y. 2023. Hugginggpt: Solving ai tasks with chatgpt and its friends in hugging face. *Advances in Neural Information Processing Systems*, 36: 38154–38180.

Soni, S.; Dudhane, A.; Debary, H.; Fiaz, M.; Munir, M. A.; Danish, M. S.; Fraccaro, P.; Watson, C. D.; Klein, L. J.; Khan, F. S.; et al. 2025. Earthdial: Turning multi-sensory earth observations to interactive dialogues. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 14303–14313.

Sun, X.; Wang, P.; Yan, Z.; Xu, F.; Wang, R.; Diao, W.; Chen, J.; Li, J.; Feng, Y.; Xu, T.; et al. 2022a. FAIR1M: A benchmark dataset for fine-grained object recognition in high-resolution remote sensing imagery. *ISPRS Journal of Photogrammetry and Remote Sensing*, 184: 116–130.

Sun, Y.; Feng, S.; Li, X.; Ye, Y.; Kang, J.; and Huang, X. 2022b. Visual grounding in remote sensing images. In *Proceedings of the 30th ACM International conference on Multimedia*, 404–412.

Tay, Y.; Phan, M. C.; Tuan, L. A.; and Hui, S. C. 2017. Learning to rank question answer pairs with holographic dual LSTM architecture. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 695–704. ACM.

Wang, J.; Wang, S.; and Zhang, Y. 2025. Deep learning on medical image analysis. *CAAI Transactions on Intelligence Technology*, 10(1): 1–35.

Wang, J.; Zhang, H.; Hong, H.; Jin, X.; He, Y.; Xue, H.; and Zhao, Z. 2023. Open-vocabulary object detection with an open corpus. In *Proceedings of the IEEE/CVF international conference on computer vision*, 6759–6769.

Wu, Z.; Chen, X.; Pan, Z.; Liu, X.; Liu, W.; Dai, D.; Gao, H.; Ma, Y.; Wu, C.; Wang, B.; et al. 2024. DeepSeek-VL2: Mixture-of-Experts Vision-Language Models for Advanced Multimodal Understanding.

Xia, G.-S.; Bai, X.; Ding, J.; Zhu, Z.; Belongie, S.; Luo, J.; Datcu, M.; Pelillo, M.; and Zhang, L. 2018. DOTA: A large-scale dataset for object detection in aerial images. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 3974–3983.

Xia, G.-S.; Hu, J.; Hu, F.; Shi, B.; Bai, X.; Zhong, Y.; Zhang, L.; and Lu, X. 2017. AID: A benchmark data set for performance evaluation of aerial scene classification. *IEEE Transactions on Geoscience and Remote Sensing*, 55(7): 3965–3981.

Xia, G.-S.; Yang, W.; Delon, J.; Gousseau, Y.; Sun, H.; and Maître, H. 2010. Structural high-resolution satellite image indexing. In *ISPRS TC VII Symposium-100 Years ISPRS*, volume 38, 298–303.

Yuan, Z.; Zhang, W.; Fu, K.; Li, X.; Deng, C.; Wang, H.; and Sun, X. 2021. Exploring a Fine-Grained Multiscale Method for Cross-Modal Remote Sensing Image Retrieval. *IEEE Transactions on Geoscience and Remote Sensing*, 60: 1–19.

Zhan, Y.; Xiong, Z.; and Yuan, Y. 2023. Rsvg: Exploring data and models for visual grounding on remote sensing data. *IEEE Transactions on Geoscience and Remote Sensing*, 61: 1–13.

Zhang, W.; Cai, M.; Zhang, T.; Zhuang, Y.; and Mao, X. 2024. Earthgpt: A universal multi-modal large language model for multi-sensor image comprehension in remote sensing domain. *IEEE Transactions on Geoscience and Remote Sensing*.

Zhang, Y.; Yuan, Y.; Feng, Y.; and Lu, X. 2019. Hierarchical and robust convolutional neural network for very high-resolution remote sensing object detection. *IEEE Transactions on Geoscience and Remote Sensing*, 57(8): 5535–5548.

Zhou, Y.; Lan, M.; Li, X.; Ke, Y.; Jiang, X.; Feng, L.; and Zhang, W. 2025. GeoGround: A Unified Large Vision-Language Model for Remote Sensing Visual Grounding. arXiv:2411.11904.

Zhu, D.; Chen, J.; Shen, X.; Li, X.; and Elhoseiny, M. 2024. MiniGPT-4: Enhancing Vision-Language Understanding with Advanced Large Language Models. In *The Twelfth International Conference on Learning Representations*.

Zhu, P.; Wen, L.; Du, D.; Bian, X.; Fan, H.; Hu, Q.; and Ling, H. 2021. Detection and tracking meet drones challenge. *IEEE transactions on pattern analysis and machine intelligence*, 44(11): 7380–7399.