ISPRS Journal of Photogrammetry and Remote Sensing 227 (2025) 539–550

Contents lists available at ScienceDirect

ISPRS Journal of Photogrammetry and Remote Sensing

journal homepage: www.elsevier.com/locate/isprsjprs

LHRS-Bot-Nova: Improved multimodal large language model for remote sensing vision-language interpretation

Zhenshi Li [a,b], Dilxat Muhtar [a], Feng Gu [a], Yanglangxing He [a], Xueliang Zhang [a,*], Pengfeng Xiao [a], Guangjun He [d], Xiaoxiang Zhu [b,c,*]

[a] Jiangsu Provincial Key Laboratory of Geographic Information Science and Technology, Key Laboratory for Land Satellite Remote Sensing Applications of Ministry of Natural Resources, School of Geography and Ocean Science, Nanjing University, Nanjing, 210023, China
[b] Data Science in Earth Observation, Technical University of Munich, Munich, 80333, Germany
[c] Munich Center for Machine Learning, Munich, 80333, Germany
[d] State Key Laboratory of Space-Ground Integrated Information Technology, Beijing, 100095, China

ARTICLE INFO

Keywords:
Remote sensing
Earth observation
Multimodal large language model
Vision-language dataset

ABSTRACT

Automatically and rapidly understanding Earth's surface is fundamental to our grasp of the living environment and informed decision-making. This underscores the need for a unified system with comprehensive capabilities in analyzing Earth's surface to address a wide range of human needs. The emergence of multimodal large language models (MLLMs) has great potential in boosting the efficiency and convenience of intelligent Earth observation. These models can engage in human-like conversations, serve as unified platforms for understanding images, follow diverse instructions, and provide insightful feedbacks. In this study, we introduce LHRS-Bot-Nova, an MLLM specialized in understanding remote sensing (RS) images, designed to expertly perform a wide range of RS understanding tasks aligned with human instructions. LHRS-Bot-Nova features an enhanced vision encoder and a novel bridge layer, enabling efficient visual compression and better language-vision alignment. To further enhance RS-oriented vision-language alignment, we propose a large-scale RS image-caption dataset, generated through feature-guided image recaptioning. Additionally, we introduce an instruction dataset specifically designed to improve spatial recognition abilities. Extensive experiments demonstrate superior performance of LHRS-Bot-Nova across various RS image understanding tasks. We also evaluate different MLLM performances in complex RS perception and instruction following using a complicated multi-choice question evaluation benchmark, providing a reliable guide for future model selection and improvement. Data, code, and models will be available at https://github.com/NJU-LHRS/LHRS-Bot.

1. Introduction

Interpreting remote sensing (RS) imagery and understanding multi-level features, object relationships, and their dynamic trends, play a significant role in various applications, such as urban sustainable development (Zhang et al., 2022; Wu et al., 2023; Sun et al., 2020), early-warning systems (Ravuri et al., 2021; Reichstein et al., 2024; Ravuri et al., 2021; Xu et al., 2024a), and earth surface processes (Qian et al., 2024; Zhu et al., 2022b). Artificial intelligence (AI) has revolutionized RS data analysis (Zhao et al., 2024; Ma et al., 2019; Reichstein et al., 2019; Zhang and Zhang, 2022; Zhu et al., 2017), and recent advancements in visual foundation models have further improved the efficiency and quality of interpretation of Earth's surface using RS data (Xiong et al., 2024; Guo et al., 2024; Hong et al., 2024; Zhu et al., 2024). However, a major drawback of visual foundation models is their need for tailored designs for specific downstream tasks, leading to fixed functions and limited generalization capabilities. Additionally, they lack the ability to interact with humans, making it difficult to fully address diverse human needs (Zhou et al., 2024).

Language, as the primary medium of human communication, plays a fundamental role in facilitating interaction with machines. Large language models (LLMs), such as ChatGPT (OpenAI, 2023a), have demonstrated remarkable conversational abilities, step-by-step reasoning skills, and the capacity to serve as general-purpose task solvers (Touvron et al., 2023; Brown et al., 2020; Zhao et al., 2023). Taking a further step toward human-level AI, multimodal large language models (MLLMs) enhance LLMs with visual perception, enabling them to see and understand the world (Fei et al., 2024; Yin et al., 2023; Achiam et al., 2023). These models have demonstrated scalability and generalizability as general-purpose assistants (Li et al., 2024c), and have already shown their strong capabilities in understanding RS data for

* Corresponding authors.
E-mail addresses: zxl@nju.edu.cn (X. Zhang), xiaoxiang.zhu@tum.de (X. Zhu).

https://doi.org/10.1016/j.isprsjprs.2025.06.003
Received 31 August 2024; Received in revised form 4 March 2025; Accepted 4 June 2025
Available online 1 July 2025
0924-2716/

real-world application (Zhang and Wang, 2024; Tan et al., 2023; Ricci et al., 2024). Developing specialized MLLMs for interpreting RS images offers several advantages: **(1) unified modeling**: MLLMs provide a versatile framework for handling a wide range of visual tasks across different granularities; **(2) human–computer interaction**: MLLMs can interpret human intent and incorporate auxiliary information through conversational interactions; **(3) reasoning**: advanced reasoning capabilities, such as chain-of-thought methods (Wei et al., 2022; Mitra et al., 2024), enable MLLMs to understand complex relations and deal with complex scenarios; and **(4) enhanced multimodal task potential**: pretrained on extensive and diverse datasets, MLLMs establish a robust baseline for addressing complex multimodal problems.

The potential of specialized MLLMs has been widely acknowledged by the research community (Moor et al., 2023; Zhou et al., 2024; Li et al., 2024e), prompting several early efforts to develop large-scale vision-language datasets and RS-specific MLLMs (Yuan et al., 2024; Muhtar et al., 2024; Kuckreja et al., 2024; Zhang et al., 2024b; Irvin et al., 2024; Pang et al., 2024a). However, we have identified three main drawbacks in current studies. **(1) Lack of high-quality, large-scale image-caption dataset:** High-quality vision-language pretraining datasets are crucial for developing robust multimodal models (Nguyen et al., 2022; Fang et al., 2022; Gadre et al., 2024; Xu et al., 2023). Several extensive RS image-caption datasets have been developed to enhance vision-language training (Wang et al., 2024b; Zhang et al., 2023; Muhtar et al., 2024). However, these datasets often suffer from noisy and uninformative captions, limited semantic richness, poor sentence diversity, and an overfocus on salient objects, which undermine effective modality alignment for RS MLLMs. **(2) Weaknesses in spatial recognition and hallucination tendencies:** We discover that current RS MLLMs exhibit low accuracy in spatial positioning and frequently produce hallucinated responses (Bai et al., 2024) when confronted with questions beyond their capabilities. **(3) Challenges in holistically evaluating MLLMs:** The ability of MLLMs to solve various visual tasks serves them as versatile multi-taskers. Although they often excel in metrics for common tasks like classification, visual question answering, and visual grounding, these metrics fall short of fully reflecting the all-round capabilities of MLLMs—particularly in recognizing complex scenes, objects, attributes, spatial relationships, and, most importantly, following human instructions.

In this study, we mitigate the above issues and propose LHRS-Bot-Nova, an improved RS-specialized MLLM for holistic interpreting RS images with human instructions (Muhtar et al., 2024). LHRS-Bot-Nova can respond to user instructions and achieve various RS interpretation tasks with state-of-the-art (SOTA) performance. To enhance RS-oriented vision-language alignment, we construct a large-scale RS image-caption dataset, LHRS-Align-Recap, by prompting an off-the-shelf multimodal captioner with RS images paired with their OpenStreetMap (OSM) features. Compared to captions generated solely by a language model using textual information (Muhtar et al., 2024), utilizing a vision-capable captioner results in richer captions with significantly enhanced language richness and improved alignment between images and captions (Table 2). It also provides more detailed descriptions, including additional recognition of geographical objects and a wider range of attributes such as location and color (Fig. 1).

To enhance the model's spatial awareness, we extend the LHRS-Instruct dataset (Muhtar et al., 2024) with conversations that primarily focus on localization and perception. Considering the necessity of a vision-centric (Tong et al., 2024) design for holistic visual understanding, we scale up the vision encoder to accommodate inputs with larger resolutions. Additionally, we propose a novel bridge layer using the MoE architecture (Jiang et al., 2024) to further extend the model's capacity, enabling lossless compression of visual information and dynamic mapping to the language domain. With enriched instructional data and an optimized modeling architecture, LHRS-Bot-Nova demonstrates exceptional spatial recognition capabilities with enhanced robustness.

Finally, we conduct a thorough evaluation of various general-purpose and RS-specific MLLMs, not only on standard RS tasks such as classification, visual question answering, and visual grounding, but also on a multiple-choice question (MCQ) evaluation benchmark, LHRS-Bench (Muhtar et al., 2024), which is designed to comprehensively assess MLLMs in the RS domain. This facilitates a holistic assessment of instruction-following abilities and other RS-specific capabilities across multiple dimensions, such as perception and spatial awareness.

The main contributions of this study are:

1. We propose a large-scale RS image-caption dataset, LHRS-Align-Recap, with high-quality captions generated through feature-guided image recaptioning. Additionally, we enlarge the instruction dataset by generating more conversations that emphasize spatial recognition and robustness.
2. We scale up the vision encoder for higher-resolution inputs and design an MoE-based bridge layer to enhance model capacity. This enables more efficient compression of visual information with limited vision tokens, thereby improving language-vision alignment performance and visual understanding.
3. We introduce a RS-specialized MLLM, LHRS-Bot-Nova, and systematically evaluate its performance across a range of tasks to assess the fundamental task-solving capabilities. Additionally, our comprehensive evaluation through an MCQ dataset provides a deeper understanding of the reliability of MLLMs as task solvers and offers valuable insights for future improvements.

2. Related works

2.1. MLLM development in the RS community

MLLMs, such as GPT-4V (OpenAI, 2023b) and Gemini (Team et al., 2023), serve as versatile assistants with advanced capabilities in visual comprehension, reasoning, and human interaction. Their success has also fueled the development of RS-specific MLLMs designed for enhanced RS image interpretation (Li et al., 2024e; Zhou et al., 2024). RSGPT (Hu et al., 2023) is the first study to incorporate LLMs for RS visual-language tasks, followed by the creation of various MLLMs (Kuckreja et al., 2024; Zhang et al., 2024b; Zhan et al., 2024) specifically trained with instruction tuning datasets generated from RS images. Beyond that, (Muhtar et al., 2024) introduced LHRS-Bot, which leverages a newly proposed large-scale RS vision-language dataset and enhances visual features with a novel vision perceiver. Pang et al. (2024a) and Pang et al. (2024b) proposed VHM, featuring enhanced self-awareness capabilities within the MLLM by introducing the first RS-specific honest dataset. Luo et al. (2024) developed SkySenseGPT, supported by a large-scale instruction tuning dataset that features complex scenes. Collectively, these studies have significantly advanced the application of MLLMs in the RS community. However, current RS MLLMs still face several challenges, including a shortage of large-scale vision-language datasets with high-quality image captions, limitations in spatial recognition, a tendency toward hallucinations, and a lack of comprehensive performance evaluation—issues that this study aims to address and improve.

2.2. Large-scale RS vision-language dataset

A diverse and extensive dataset is the most crucial element for effective vision-language model training. Additionally, the quality of captions plays a significant role in aligning the vision and language modalities in MLLMs (Nguyen et al., 2024; Chen et al., 2023b; Li et al., 2024d). Several large-scale RS image-caption datasets have been introduced to improve multimodal alignment in RS scenarios. RS5M (Zhang et al., 2023) is the first extensive RS image-caption dataset, but it is assembled from web-crawled data, which is often noisy and lacks informativeness (Nguyen et al., 2024). SkyScript (Wang et al., 2024b) comprises accurately paired images and captions, but the captions are generated using simple rules, resulting in a lack of semantic richness, which is essential for effectively training MLLMs (Wang et al., 2024b). To address these issues, LHRS-Align (Muhtar et al., 2024) generates captions based on OSM features by employing an LLM. However, these captions lack sentence diversity and may focus on salient objects, which is sub-optimal for modality alignment (Chen et al., 2023b). In this study, we explore feature-guided image recaptioning (Chen et al., 2023b; Li et al., 2024d) to enhance the quality of vision-language pre-training datasets. A study similar to ours is VHM (Pang et al., 2024a), in which text descriptions are generated from images using an MLLM. However, there are three key differences between our approaches: VHM uses the closed-source Gemini model, while we use an open-source model, making our approach cost-free and more sustainable. VHM relies solely on image features for generating captions, whereas we combine image features with OSM attributes as text guidance, providing richer and more accurate descriptions. VHM uses open-source RS datasets, which are limited in scope, while ours, based on global OSM features, is more scalable and covers more land cover types.

3. Dataset

3.1. LHRS-align-recap: feature-guided image recaptioning for multimodal alignment

High-quality vision-language alignment datasets play a pivotal role for training a robust multimodal model (Nguyen et al., 2022; Fang et al., 2022; Gadre et al., 2024; Xu et al., 2023). To enhance vision-language alignment for RS-specific multimodal models, LHRS-Align (Muhtar et al., 2024) was proposed by pairing RS images with OSM features, which are then used for generating captions with an LLM. Though effective for improving vision-language alignment, the captions in LHRS-Align are often concise and grammatically monotonous, and focus on key objects constrained by OSM features, as shown in Fig. 1, leading to suboptimal alignment across modalities (Chen et al., 2023b).

To further enhance caption quality for more robust multimodal alignment, we leverage an MLLM for recaptioning and propose a new large-scale RS vision-language dataset, LHRS-Align-Recap. Specifically, we prompt the Share-Captioner (Chen et al., 2023b) to generate captions based on RS images and their corresponding OSM features from the LHRS-Align dataset (Muhtar et al., 2024), using the prompt design shown in Table 1. In contrast to LHRS-Align (Muhtar et al., 2024), where captions are generated by a blind LLM, LHRS-Align-Recap is produced with an MLLM that can actually perceive the image. This allows it to not only comprehend image features but also interpret textual guidance, utilizing OSM attributes, to ensure the final description is both comprehensive and accurate. As a result, it produces highly descriptive captions with greater detail and diverse sentence structures, as illustrated by the examples in Fig. 1.

To gain a deeper understanding of the improvement, we conduct a statistical analysis of the original and new captions, focusing on two key aspects: the inherent distribution of the captions and the alignment quality of the captions. From the first perspective, we calculate the number of unique words and unique trigrams for both versions of the captions. These metrics gauge vocabulary richness and the structure and repetitive patterns, respectively. As shown in Table 2, the number of unique words and trigrams in LHRS-Align-Recap is nearly twice that of LHRS-Align, indicating increased vocabulary diversity and more varied sentence structures and phrase constructions. Additionally, the new captions have an average sentence length of 150 words, which is nearly five times that of the original captions, with a more even distribution, as shown in Fig. 2. In terms of alignment quality, we use the CLIP score computed from LongCLIP (Zhang et al., 2024a) to assess the semantic alignment between captions and images. The higher CLIP score computed from the LHRS-Align-Recap dataset, as shown in Fig. 2 and Table 2, demonstrates its greater effectiveness for vision-language alignment (Nguyen et al., 2024). A detailed quantitative validation of the improved dataset will be discussed in Section 5.5.

Table 1
Prompt for generating captions with Share-Captioner based on RS images and OSM features.

> Given the following image and its associated key-value features, generate a concise and descriptive caption that captures the essence of the image. The caption should reflect the relationships, context, and any notable details highlighted by the features. Ensure that the caption is coherent and informative, making use of the provided tags to enhance accuracy and detail.
> `{image}`
> `{Key-Value Features}`

Table 2
Caption improvements in LHRS-Align-Recap compared to the original LHRS-Align, highlighting greater vocabulary richness, more varied sentence structures, and stronger alignment with images.

Dataset	No. of unique words	No. of unique trigrams	Average sentence length	Average LongCLIP score
LHRS-Align	8436	1.12×10^6	31	70.81
LHRS-Align-Recap	15 345	2.64×10^6	150	88.12

3.2. Instruction tuning datasets for training LHRS-Bot-Nova

Visual instruction fine-tuning (Liu et al., 2024b) plays a significant role in MLLM training by enabling LLMs to better understand visual features and follow human instruction (Xu et al., 2024b). This process enhances the model's capabilities and controllability, allowing it to generalize to diverse tasks and provide feedbacks that align with human preferences (Li et al., 2023a; Zhang et al., 2023b). We introduce four datasets used for instruction tuning LHRS-Bot-Nova as below, including a multi-task dataset, LHRS-Instruct, LRV-Instruct, and the proposed LHRS-Instruct-Plus (Table 3).

We utilize the multi-task instruction dataset and the LHRS-Instruct dataset proposed in Muhtar et al. (2024) to enhance task-solving and complex understanding capabilities. The former was constructed by combining various public RS datasets with manually created instruction templates, and the latter was created by prompting LLMs to create complex conversations using selected samples from RS caption datasets.

To further improve the ability of LHRS-Bot-Nova for understanding spatial relationships, we construct a novel instruction dataset, LHRS-Instruct-Plus. Specifically, the proposed dataset is generated by prompting GPT-4V to generate various conversations from the two RS object detection datasets, DOTAv2 (Ding et al., 2021) and FAIR1M (Sun et al., 2022a). The prompt we used is shown in Table 4. The generated conversations primarily includes object identification, as well as other tasks such as object counting, image description, and visual reasoning.

Furthermore, recent studies have shown that current MLLMs tend to answer "Yes" to any given instruction, even when the correct response should be "No", a phenomenon known as hallucination (Li et al., 2023b; Bai et al., 2024; Liu et al., 2023a). One reason for this issue is the imbalance in instruction datasets, where positive instructions predominate and the importance of negative samples is overlooked. To mitigate this problem, we integrate the LRV-Instruct dataset (Liu et al., 2023a), which contains both positive and negative instructions, into our instruction datasets, for more robust visual instruction tuning. This makes the trained model more reliable and enhances its ability to accurately differentiate between affirmative and negative responses, thereby reducing the occurrence of hallucinations.

In general, all the instruction tuning datasets for training our LHRS-Bot-Nova are listed in Table 3, and some examples are shown in Fig. 3.

Fig. 1. Four examples from LHRS-Align and LHRS-Align-Recap. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

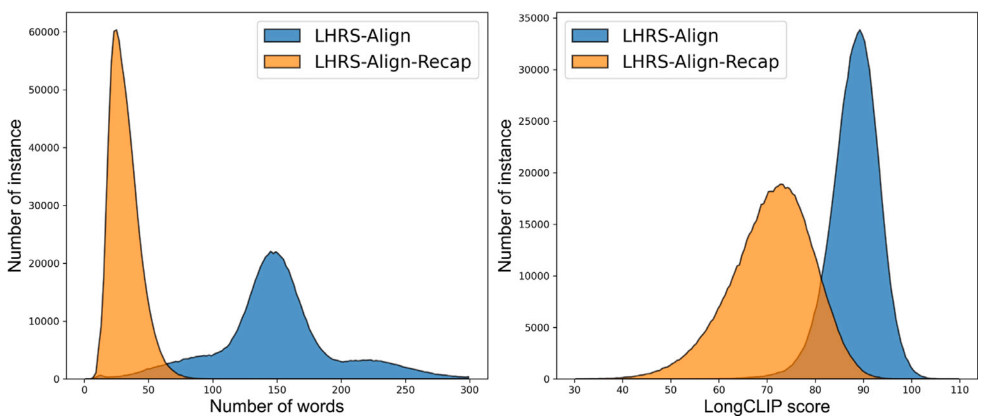


Fig. 2. Caption length (left) and image-caption LongCLIP score (right) distributions for the LHRS-Align and the proposed LHRS-Align-Recap dataset.

Table 3
Instruction tuning datasets for training LHRS-Bot-Nova.

Dataset	Original dataset type	Original dataset	Instance num.	Instruction task
Multitask	VQA	RSVQA-HR (Lobry et al., 2020) RSVQA-LR (Lobry et al., 2020)	213 160 11 440	VQA with consize answers
	Classification	UCM (Yang and Newsam, 2010) fMoW (Christie et al., 2018) METER-ML (Zhu et al., 2022a)	2100 5352 1400	Classification
	Visual grounding	RSVG (Sun et al., 2022b) DIOR-RSVG (Zhan et al., 2023)	2428 14 030	Visual grounding
	Image captioning	RSICD (Lu et al., 2018)	1000	Briefly image captioning
LHRS-Instruct	Image captioning	NWPU (Cheng et al., 2022) RSITMD (Yuan et al., 2022) LHRS-Align (Muhtar et al., 2024)	111 755 12 200 29 671	Conversation Conversation Conversation, Detailed description, Visual reasoning
LHRS-Instruct-Plus	Object detection	DOTAv2 (Ding et al., 2021) FAIR1M (Sun et al., 2022a)	67 143 253 446	Conversation, Object detection Conversation, Object detection
LRV-Instruct	Instruction tuning	LRV-Instruct (Liu et al., 2023a)	140 974	Robust Visual Instruction with negative samples

4. Methodology

LHRS-Bot-Nova incorporates an enhanced vision encoder and a new vision perceiver with an MoE structure for improved vision-language alignment. In this section, we delve into each component of LHRS-Bot-Nova, detailing how the enhanced architecture achieves better vision-language alignment. Subsequently, we explain the curriculum training strategy used for training LHRS-Bot-Nova.

4.1. Model architecture

LHRS-Bot-Nova is primarily composed of three components: a vision encoder, a vision perceiver, and a foundational LLM. The overall architecture of LHRS-Bot-Nova is presented in Fig. 4.

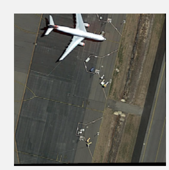
Vision Encoder. Recent studies have highlighted that increasing the input image resolution enhances the visual understanding ability of MLLMs (Tong et al., 2024; Liu et al., 2024a; Li et al., 2024b). Therefore, we adopt SigLIP-L/14 (Zhai et al., 2023) with an input resolution of

Fig. 3. Six examples of various instruction tasks from the instruction datasets used to train LHRS-Bot-Nova.

Table 4
Prompt for generating LHRS-Instruct-Plus with GPT-4V based on RS object detection datasets.

System Message:
You are an AI visual assistant that can design a conversation for analyzing images. You will be presented with an image along with the category of the main object and the coordinates of its bounding box, given as (x1, y1, x2, y2), with values ranging from 0 to 1. These coordinates correspond to the top-left x, top-left y, bottom-right x, and bottom-right y positions, respectively. Your task is to design conversations about the image based on the object's categories and their locations. The dialogs should primarily focus on the number of objects, their locations, and the spatial relationships between them. When designing conversation, use the information from the coordinates naturally, without mentioning that the source is the given bounding box.

In-Context Example:
User:
airplane [0.399,0.163,0.452,0.293]



Assistant:
Questions:
How many airplanes can you see?
Answer:
I can see an airplane in the image.
Questions:
Where is the airplane positioned in the image?
Answer:
The airplane is mostly in the middle of the image but slightly tilted toward the right. Its tail end is closer to the center, while the nose is pointing slightly upwards and to the right.
Questions:
[DET] Please give me the exact position of the airplane.
Answer:
<bbox>[0.399,0.163,0.452,0.293]</bbox>

Table 5
Hyperparameter settings for each stage.

	Stage1	Stage2	Stage3
Learning rate	0.0002	0.0001	0.0001
Global batch size	128	64	64
Weight decay	0	0.01	0.01
β_1		0.9	
β_2		0.95	
Gradient norm		1.0	
Scheduler		Cosine	
Warmup steps	300	100	0

336 × 336 as the vision encoder to extract more detailed visual signals. Additionally, we follow the strategy in Muhtar et al. (2024) to extract multi-level visual information, providing additional visual supervision for more data-efficient vision-language alignment.

Vision Perceiver. Considering the additional computational and memory overhead associated with the extra vision tokens from multi-layer visual signals, we follow Muhtar et al. (2024) to summarize different layers of visual signals using learnable query-based cross-attention. Additionally, we adopt a decreasing query allocation strategy to manage the higher redundancy in deeper levels of the vision hidden states within the vision encoder (Bolya et al., 2022). However, while this design effectively compresses vision tokens by reducing their number, it may lead to a loss of visual details (Tong et al., 2024). Increasing

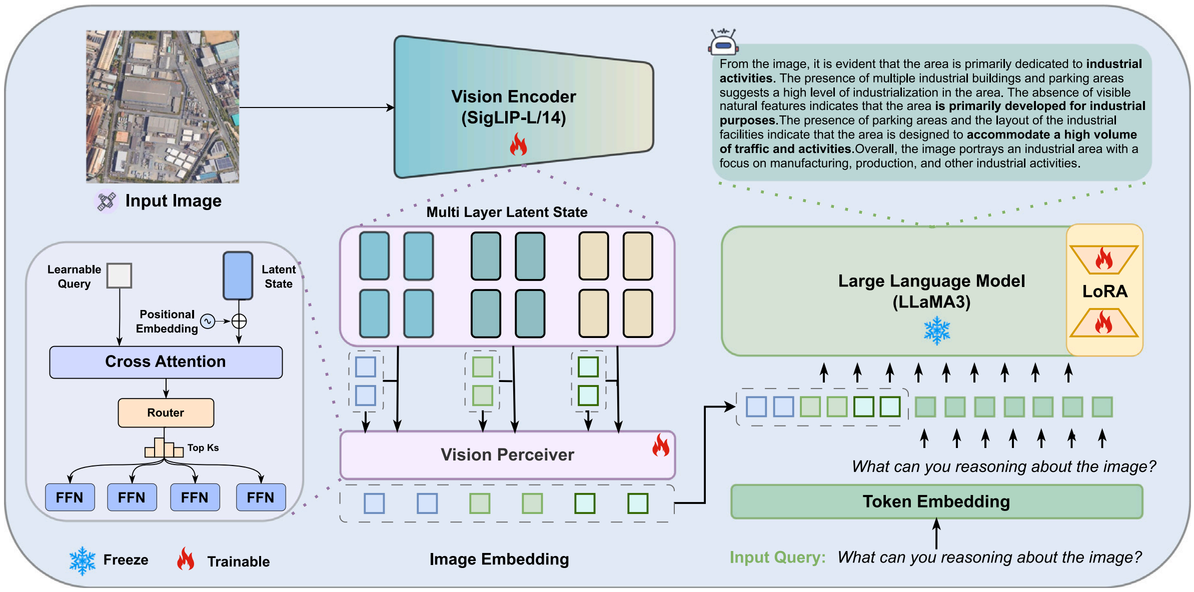


Fig. 4. Architecture of LHRS-Bot-Nova. LHRS-Bot-Nova employs learnable queries in conjunction with an MoE vision perceiver to summarize multi-level visual representations, which are then concatenated with language token embeddings as input to the LLM.

the token count is a straightforward way to preserve detail information, but it is not computationally efficient. Therefore, inspired by the observation that the feed-forward network (FFN) layer acts as network memory (Geva et al., 2021), we incorporate MoE architecture (Jiang et al., 2024) into each layer of the FFN in the Vision Perceiver to expand the model's memory capacity, allowing for the extraction of more detailed visual information even with fewer vision tokens. Compared to increasing the token count, the sparse activation characteristic of MoE introduces only a modest increase in computational and memory overhead, which can be further optimized through parallelization.

Specifically, given the learnable query $\mathbf{Q}^i \in \mathbb{R}^{n^i \times d}$ and the vision token $\mathbf{X}^i \in \mathbb{R}^{L \times d}$, where the superscript i denotes the ith level (with $i \in 1, 2, 3$ typically corresponding to low, medium, and deep levels), L denotes the number of tokens, and d is the hidden dimension, we first summarize $\mathbf{X}^i$ using $\mathbf{Q}^i$ through cross-attention:

$$\mathbf{h}^i = \mathrm{Softmax}\left(\frac{\mathbf{Q}^i(\mathbf{W}_k\mathbf{X}^i + p)^T}{\sqrt{d}}\right)(\mathbf{W}_v\mathbf{X}^i + p) \in \mathbb{R}^{n^i \times d}, \tag{1}$$

where $\mathbf{W}_k, \mathbf{W}_v \in \mathbb{R}^{d \times d}$ are learnable parameters, p denotes sinusoidal positional embedding, and $\mathbf{h}_i$ represents the summarized vision tokens for the ith level. Then, we concatenate all levels of summarized results, $\mathbf{h} := [\mathbf{h}^1; \mathbf{h}^2; \mathbf{h}^3] \in \mathbb{R}^{(n^1+n^2+n^3) \times d}$. For each token $\mathbf{h}_t$ in the concatenated result, we compute the output $\mathbf{O}_t$ of the MoE-FFN layer as follows:

$$\mathbf{O}_t = \mathbf{h}_t + \sum_{j=1}^{N_e} g_{j,t}\mathrm{FFN}_j(\mathbf{h}_t) \tag{2}$$

$$g_{j,t} = \begin{cases} s_{j,t}, & s_{j,t} \in \mathrm{Top}K(\{s_{k,t} | 1 \le k \le N_e\}, K), \\ 0, & \text{otherwise}, \end{cases} \tag{3}$$

$$s_{j,t} = \mathrm{Softmax}_j(\mathbf{h}_t^T \mathbf{W}_{\mathrm{router}}), \tag{4}$$

where N_e is the number of experts, $\mathrm{FFN}_j(\cdot)$ denotes the jth expert, K represents the number of activated routed experts for each token, $g_{j,t}$ is the gate value for the jth expert, $s_{j,t}$ is the token-to-expert affinity, $\mathbf{W}_{\mathrm{router}}$ is the learnable parameter for the expert router, and $\mathrm{Top}K(\cdot, K)$ denotes the set comprising the K highest scores among the affinity scores calculated for the tth token across all routed experts. With this MoE architecture in the FFN layer and the per-token sparse dynamic routing strategy, we expand the model's memory capacity without introducing additional inference activation, allowing for the extraction and retention of more detailed visual information.

Large Language Model. We employ the improved LLaMA3-8B (Dubey et al., 2024) architecture as the central "brain" of LHRS-Bot-Nova, enabling it to interpret various signals from vision and language to respond to given instructions.

4.2. Training strategy

We follow the three-stage curriculum learning strategy introduced by Muhtar et al. (2024), which includes pre-training, multi-task instruction fine-tuning, and supervised fine-tuning stages.

In the pre-training stage, we use the LHRS-Bot-Recap dataset to train the vision encoder and vision perceiver, mapping multi-level visual signals into the language domain. We unfreeze the vision encoder (only at this stage) to enhance its ability to extract fine-grained features.

During the multi-task instruction fine-tuning stage, we unfreeze the LLM with low-rank adapters (LoRA) (Hu et al., 2021) and fine-tune both the vision perceiver and LoRA using the multi-task instruction dataset to improve the multimodal task-solving capabilities of LHRS-Bot-Nova.

Finally, in the supervised fine-tuning stage, we use all the instruction data from the LHRS-Instruct, LHRS-Instruct-Plus, and LRV-Instruct (Liu et al., 2023a) datasets to further train the vision perceiver and LoRA adapter, enhancing the conversational and reasoning capabilities of LHRS-Bot-Nova.

5. Experiments

5.1. Experimental setup

Evaluation benchmarks. We evaluate our model in various RS image understanding benchmarks detailed as follows. Noted that we also conducted tests on image captioning datasets. However, due to the high complexity of RS scenes, ground-truth captions of current datasets often appear simplistic and insufficiently detailed, making it difficult to accurately evaluate the quality of the model's output. We have therefore excluded them from this paper, as the results are not suitable as a reference.

(1) Scene classification. We use the test sets of seven datasets: AID (Xia et al., 2017), WHU-RS19 (Dai and Yang, 2011), SIRI-WHU (Zhu et al., 2016), EuroSAT (Helber et al., 2019), NWPU (Cheng et al., 2017), METER-ML (Zhu et al., 2022a), and fMoW (Christie et al., 2018). The first four datasets are evaluated in a zero-shot setting.

Table 6
Accuracies (%) of different MLLMs on various scene classification datasets.

Method	AID	WHU-RS19	SIRI-WHU	EuroSAT	NWPU	METER-ML	fMoW	Avg.
LLaVA-1.5	31.10	54.55	17.71	26.12	34.96	21.73	11.43	28.23
MiniGPTv2	32.96	64.80	35.46	38.56	28.15	14.29	5.20	31.35
InstructBLIP	29.50	36.76	18.20	20.25	34.01	14.42	6.71	22.84
mPLUG-OWL2	48.79	72.66	54.83	33.29	46.58	36.27	17.85	44.32
QWen-VL-Chat	55.30	72.25	54.58	26.42	42.73	38.77	6.89	42.42
InternLM-XComposer	51.61	72.89	46.83	39.70	47.51	40.21	11.28	44,29
LHRS-Bot	**91.26**	93.17	62.66	51.40	83.94	69.81	56.56	71.83
LHRS-Bot-Nova	88.32	**95.63**	**74.75**	**63.54**	**86.80**	**70.05**	**57.11**	**76.60**

(2) Visual question answering (VQA). We utilize two datasets: the test sets of RSVQA-HR and RSVQA-LR (Lobry et al., 2020).

(3) Visual grounding. We use two datasets: the test sets of RSVG (Sun et al., 2022b) and DIOR-RSVG (Zhan et al., 2023).

(4) Benchmarks for RS MLLMs. LHRS-Bench (Muhtar et al., 2024) with MCQs is used for systematic evaluation of MLLMs in RS image understanding.

Baselines. We evaluate LHRS-Bot-Nova against the 7B variants of several powerful open-source general domain MLLMs, including LLaVA-1.5 (Liu et al., 2024a), MiniGPTv2 (Chen et al., 2023a), InstructBLIP (Dai et al., 2023), mPLUG-Owl2 (Ye et al., 2024), QWen-VL-Chat (Bai et al., 2023), and InternLM-Xcomposer (Zhang et al., 2023a), across various tasks. For comparisons with RS MLLMs, we use the accuracies reported in the papers whenever possible, including GeoChat (Kuckreja et al., 2024), SkyEyeGPT (Zhan et al., 2024), H2RSVLM (Pang et al., 2024b), and SkySenseGPT (Luo et al., 2024).

Implementation details. We extract the latent states at layers $\{N_L/3, 2N_L/3, N_L-1\}$ for summarization through the MoE vision and LHRS-Bot (Muhtar et al., 2024) perceiver, where N_L denotes the number of layers in the vision encoder. The allocated queries for each layer are {112, 96, 64}. The vision perceiver implementation comprises six layers of cross-attention and FFN, with each layer containing four FFN experts, and the number of activated experts K in Eq. (3) is set to 2. We apply the LoRA module to every linear layer of the LLM, setting the rank r and the scale factor α for LoRA to 128 and 256, respectively. Additionally, we introduce task identifiers [CLS], [CONCISE], and [DET] for classification, concise vision-language answering, and visual grounding tasks, respectively. All 3 stages of training were done with AdamW optimizer on 8 × H100 GPUs for 1 epoch, the hyperparameters used at each stage are presented in 5.

5.2. Quantitative results on RS image understanding

Scene classification. We evaluate the scene classification accuracy of LHRS-Bot-Nova compared with other open-source MLLMs, highlighting its broad knowledge in recognizing geographical features. As shown in Table 6, the classification accuracy of LHRS-Bot-Nova surpasses other general MLLMs by a significant margin, thanks to the RS domain-specific training. Notably, compared to LHRS-Bot, the classification performance has improved across nearly all datasets, with a significant overall accuracy increase of up to 4.77%, demonstrating the powerful scene understanding ability of LHRS-Bot-Nova. Since the AID, WHU-RS19, SIRI-WHU, and EuroSAT datasets were entirely absent from the multi-task training data, the accuracies achieved on these datasets reflect a zero-shot setting, demonstrating the strong generalization capability of LHRS-Bot-Nova. Overall, the high classification accuracy highlights the effectiveness of optimizing the large-scale vision-language alignment dataset. The remarkable improvement in performance demonstrates that our proposed LHRS-Align-Recap dataset, with its precise image-text alignment and rich, detailed scene descriptions, enables the model to gain a deeper and more comprehensive understanding of geographical knowledge in RS scenes.

Visual question answering. We report the VQA results of LHRS-Bot-Nova on two RSVQA datasets by comparing with other general domain MLLMs and RS domain MLLMs in Table 7. It can be seen that the VQA results of LHRS-Bot-Nova surpass those of other general domain MLLMs by a significant margin. When compared to other RS-specific MLLMs, our model achieves comparable results on the RSVQA-LR data and demonstrates a significant advantage on the RSVQA-HR data. Overall, LHRS-Bot-Nova achieves the highest VQA accuracy, which slightly outperforms LHRS-Bot.

Visual grounding. The visual grounding accuracies for the RSVG and DIOR-RSVG datasets, using accuracy@0.5 as the evaluation metric, are presented in Table 8. In comparison with other MLLMs, LHRS-Bot-Nova achieves the highest accuracies on both datasets. Compared to LHRS-Bot, our improved model shows significantly enhanced object interpretation and localization, with a 6.58% increase in average accuracy. This improvement highlights the effectiveness of our data enhancements, particularly the instruction tuning dataset, which equips the model with specialized spatial awareness for RS objects. Additionally, the more powerful MoE-based vision encoder plays a crucial role in enabling fine-grained feature extraction.

5.3. Evaluation on LHRS-bench

Evaluation method. The evaluation of LLMs and MLLMs can be broadly categorized into generation-based and multiple-choice-based approaches. Generation-based methods involve scoring open-ended answers, typically requiring human or LLM for assessments (Zheng et al., 2024), which can introduce subjectivity. Consequently, many benchmarks employ MCQs to assess LLM and MLLM capabilities (Hendrycks et al., 2021; Zhang et al., 2024c; Liu et al., 2024c; Li et al., 2024a), which can quantitatively evaluate the model with an objective accuracy metric. In the RS field, MCQ benchmarks have also been introduced for RS MLLM evaluation (Muhtar et al., 2024; Luo et al., 2024). In this study, we thoroughly explore how to effectively evaluate RS MLLMs using MCQ benchmarks, with the LHRS-Bench (Muhtar et al., 2024) dataset as example.

Choice assignment. When directly prompting the MLLM to answer the correct choice, the model may sometimes ignore the instruction and output the full context of the candidate choices (Liu et al., 2023b). To address this, Liu et al. (2023b) applies an exact matching approach between the model's output and the correct choice. However, this method often requires additional validation of ChatGPT-based matching because exact matching frequently fails. Muhtar et al. (2024) employs substring matching, considering an answer correct if the correct context is a substring of the model's output. However, this strategy can lead to mismatches: for example, if the correct context is “industrial”, an answer like “residential but not industrial” might be mistakenly judged as correct. To resolve this, we use a simple approach: the prompt asks the model to “Only answer with the letter corresponding to the given choices, such as A., B., etc”., and the answer is considered correct only if it strictly matches the expected letter.[1] This metric can also evaluate the model's ability to follow instructions since a robust model should understand that the user is only interested in the correct letter.

[1] Interestingly, in practice, we find that all the models used for comparison can follow this simple prompt except GPT-4o-mini. Therefore, we use one-shot in-context learning to prompt GPT-4o-mini for better ability comparison.

Table 7
Accuracies (%) of various general-domain and RS-specific MLLMs on visual question answering task in the RSVQA dataset.

Method	RSVQA-LR				RSVQA-HR			Overall
	Rural/Urban	Presence	Compare	Avg.	Presence	Compare	Avg.	Avg.
LLaVA-1.5	59.22	73.16	65.19	65.86	48.96	59.02	53.99	61.11
MiniGPTv2	60.02	51.64	67.64	59.77	68.34	64.71	66.53	62.47
InstructBLIP	62.62	48.83	65.92	59.12	62.63	62.90	62.77	60.58
mPLUG-Owl2	57.99	74.04	63.69	65.24	47.60	58.47	53.04	60.36
QWen-VL-Chat	62.00	47.65	66.54	58.73	61.75	65.98	63.87	60.78
InternLM-XCompose	59.00	66.74	52.91	59.55	67.79	66.62	67.21	62.61
GeoChat	91.09	90.33	**94.00**	91.81	58.45	83.19	70.82	83.41
SkyEyeGPT	88.93	88.63	75.00	84.16	80.00	80.13	82.56	82.54
H2RSVLM	88.00	89.58	89.79	89.12	65.00	83.70	74.35	83.21
SkySenseGPT	**95.00**	**91.07**	92.00	**92.69**	69.14	84.14	76.64	86.27
LHRS-Bot	89.07	88.51	90.00	89.19	**92.57**	**92.53**	**92.55**	90.54
LHRS-Bot-Nova	89.11	89.00	90.71	89.61	91.68	92.44	92.06	**90.59**

Table 8
Comparison of different MLLMs on visual grounding with the evaluation metric of accuracy@0.5 (%), where a prediction is correct if the IoU between the predicted and ground-truth bounding boxes exceeds 0.5.

Method	RSVG	DIOR-RSVG	Avg.
QWen-VL-Chat	44.76	80.65	62.71
MiniGPTv2	46.64	85.99	66.32
SkyEyeGPT	70.50	88.59	79.55
LHRS-Bot	73.45	88.10	80.78
LHRS-Bot-Nova	**81.85**	**92.87**	**87.36**

Robust evaluation. Recent studies reveal a phenomenon of selection bias and random guessing when evaluation with MCQs (Myrzakhan et al., 2024; Wang et al., 2024a; Robinson et al., 2021), where models tend to favor certain options or provide random answers when unable to solve a question. This leads to unreliable evaluation accuracies. To address this, techniques such as option reordering, increasing or reducing options, and altering them have been proposed (Wang et al., 2024a). In our experiments, we employ CircularEval (Liu et al., 2023b, 2024c), which feeds the same question to an MLLM multiple times while rotating the correct answer among the options and checks if the model consistently selects the correct answer across all attempts. We found that using this strategy is crucial for enhancing the robustness of the MCQ evaluation for RS MLLMs: without it, the accuracy on LHRS-Bench of various MLLMs could increase by around 20%, misleading users with an inaccurate assessment of the models' abilities.

Results. Our evaluation results on the LHRS-Bench dataset are presented in Table 9, which clearly demonstrates that LHRS-Bot-Nova stands out as the best-performing approach overall. LHRS-Bot-Nova achieves the highest overall accuracy of 34.93%, a significant margin above all other methods. It not only delivers top performance in areas like identity and reasoning but also maintains a balanced distribution of strong scores across multiple dimensions. Notably, it even outperforms the closed-source models, GPT-4o-mini and Claude-3, highlighting the importance of RS-specific training for interpreting RS images. Nearly all the MLLMs, except GPT-4o-mini, struggle to recognize the resolution of RS images, which is understandable given the lack of relevant datasets for training. As a RS chatbot, while LHRS-Bot-Nova demonstrates strong effectiveness in RS image understanding, these results also highlight promising directions for enhancing the general recognition capabilities of RS MLLMs, particularly in areas such as orientation, object counting, and other complex tasks.

5.4. Multimodal dialog in RS understanding

We present conversation examples to qualitatively demonstrate LHRS-Bot-Nova performance in RS image interpretation. As shown in Fig. 5, LHRS-Bot-Nova can clearly and thoroughly describe RS image scenes, showcasing its excellent recognition of RS features. It can engage in conversation with users, following instructions for tasks like object localization and reasoning. More importantly, it does not simply respond blindly to user queries. Thanks to the utilization of the more balanced instruction data with negative samples, it can assess the validity of user instructions and is capable of providing negative responses or refusing to answer, demonstrating high reliability.

5.5. Ablation analysis

Effectiveness of improved caption quality.
To evaluate the effectiveness of data recaptioning, we visualize the loss curves for pretraining using both the original and recaptioned datasets, as shown in Fig. 6. It clearly shows that the model trained on LHRS-Align-Recap exhibits a lower loss during pretraining, which demonstrates improved convergence and suggests better vision-language alignment. This indicates that recaptioning effectively enhances the model's ability to understand and associate visual and textual modalities.

Furthermore, we compare the performance of models trained with various pretraining datasets, including the original LHRS-Align dataset, the recaptioned LHRS-Align-Recap dataset, and another SOTA RS image-caption dataset, VHM-Pretrain (Pang et al., 2024a), as summarized in Table 10. Each dataset is used separately for pretraining, while all other experimental conditions remain consistent. For different tasks, we calculate the mean accuracy across all datasets outlined in Section 5.1.

The ablation results in Table 10 demonstrate that model performance consistently improves across all tasks when pre-trained with the proposed LHRS-Align-Recap dataset compared to LHRS-Align, highlighting the effectiveness and importance of generating better captions. The most significant improvement is observed in scene classification accuracy, with a 20.39% increase, highlighting a substantial enhancement in RS scene recognition capabilities. This improvement can be attributed to the more precise image-caption alignment and the inclusion of more detailed image information descriptions. Notably, the model trained with LHRS-Align-Recap achieves a 9.1% accuracy improvement in the visual grounding task compared to using LHRS-Align. This improvement is not only due to more precise land cover identification but also demonstrates that, unlike using an LLM for caption generation (e.g., LHRS-Align), leveraging an MLLM capable of directly interpreting images allows for a more accurate capture of spatial information about geographical objects. This data enhancement equips the model with accurate object spatial recognition capabilities even during the pretraining stage.

Furthermore, when compared to the VHM-Pretrain dataset, despite LHRS-Align-Recap having a slightly smaller quantity (1.15M) than VHM-Pretrain (1.4M), LHRS-Align-Recap outperforms VHM-Pretrain across all tasks. This demonstrates the advantages of our dataset's high quality and richness in enhancing multimodal model training.

Effectiveness of various model and training configurations.

Table 9
Comparison results on the LHRS-Bench dataset with different open-source and closed-source MLLMs (unit: %).

Method	Identity	Color	Orientation	Shape	Area	Resolution	Modality	Location	Distance	Quantity	Reasoning	OA
Qwen-VL-Chat	25.87	17.70	10.26	**45.95**	17.33	0.00	0.00	16.67	9.09	8.76	43.48	24.93
LLaVA-1.5	29.34	21.24	**12.82**	27.03	24.00	0.00	4.35	25.98	**22.73**	**21.90**	58.70	28.26
InternLM-XComposer	5.05	6.19	2.56	18.92	1.33	0.00	0.00	4.41	4.55	1.46	6.52	4.64
GeoChat	20.66	9.73	7.69	32.43	14.67	0.00	13.04	12.25	9.09	9.49	45.65	19.86
GPT-4o-mini	30.13	19.47	7.69	43.24	**29.33**	**28.57**	21.74	**29.90**	13.64	8.03	39.13	30.00
Claude-3-Opus	23.82	18.58	2.56	43.24	18.67	4.76	17.39	26.47	4.55	6.57	30.43	23.19
LHRS-Bot-Nova	**36.12**	**27.43**	5.13	40.54	25.33	0.00	**26.09**	26.96	18.18	8.76	**63.04**	**34.93**

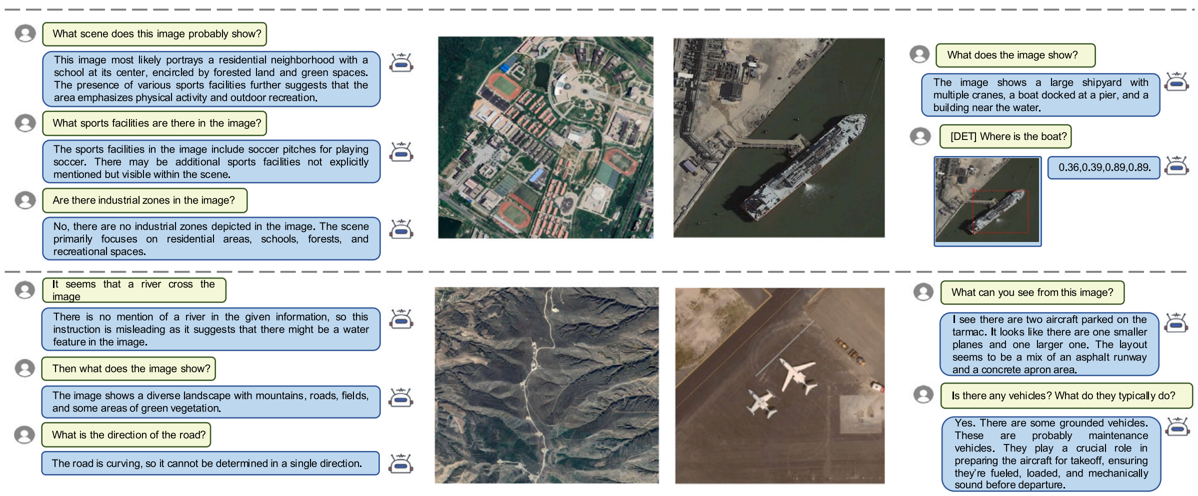


Fig. 5. Four conversation examples between user and LHRS-Bot-Nova.

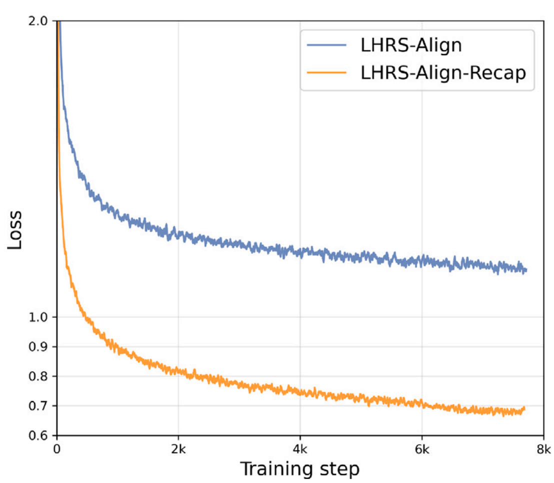


Fig. 6. Loss curve for different image-text pre-training datasets at stage 1 training.

Table 10
Accuracies (%) across various tasks for models trained with different pretraining datasets.

Dataset	Classification	VQA	Visual grounding
LHRS-Align	56.21	90.30	78.26
VHM-Pretrain	71.15	79.75	83.38
LHRS-Align-Recap	**76.60**	**90.59**	**87.36**

To verify the effectiveness of our optimized training strategy and model design, ablation experiments are conducted as shown in Table 11. Considering the high time cost of large model training, we only perform two-stage training during ablation experiments, and LLaMA2-7B-Instruct is used as the base LLM. First, we compare the performance of freezing versus unfreezing the vision encoder during the pretraining stage, as shown in the first and second rows of Table 11. It is observed that unfreezing the vision encoder enhances performance on fine-grained perception tasks, as demonstrated by the improvement in visual grounding, which aligns with recent findings (Tong et al., 2024). To mitigate the loss of detailed information due to query-based vision token compression, we employed the MoE architecture to enhance model capacity, allowing it to capture more image details while keeping low computational costs. The results in the second and third rows of Table 11 show that the MoE Vision Perceiver consistently outperforms the vanilla Vision Perceiver across all tasks, with a notable improvement of 4.14% on the classification task and 1.52% on visual grounding. This approach enhances the model's overall performance, improving its ability to capture both global and fine-grained details. Additionally, we enhanced fine-grained visual feature extraction by increasing the input image resolution from 224×224 to 336×336 and raising the query count at each level, leading to further performance gains, as shown in the last row of Table 11.

6. Conclusion

We propose LHRS-Bot-Nova, which integrates visual signals with language expression to achieve unified RS image interpretation and understanding under human instructions. With higher information density and quality in the synthetic image-caption alignment data, the improved instruction data targeted at spatial reasoning, and the vision-centric architecture design, LHRS-Bot-Nova exhibits strong performance across tasks such as scene classification, visual grounding, and question answering, while also surpassing other MLLMs in general RS interpretation. Additionally, the results of our systematic evaluation across different models and tasks provide a reliable reference for future model selection and enhancement.

Despite the impressive performance of LHRS-Bot-Nova, it still has limitations, such as insufficient faithfulness to fully avoid hallucination. We believe that designing a more rigorous alignment data curation pipeline, combined with improved training strategies like preference

Table 11
Performance comparison with different designs.

Method	Unfreeze vision encoder	Image resolution	Classification	VQA	Visual grounding
Vision Perceiver	✗	224 × 224	65.43	90.14	79.90
Vision Perceiver	✓	224 × 224	64.29	90.07	80.79
MoE Vision Perceiver	✓	224 × 224	68.43	90.22	82.31
MoE Vision Perceiver	✓	336 × 336	**71.65**	**90.27**	**84.79**

alignment, can further enhance the performance of MLLMs in interpreting RS images. Moreover, the current model is too large, highlighting the need for developing more lightweight models to meet the demands of real-time RS monitoring. Additionally, the current model is restricted to handling only RGB images. Developing multispectral MLLMs holds significant potential for improving the accuracy and versatility of RS image interpretation.

CRediT authorship contribution statement

Zhenshi Li: Writing – review & editing, Writing – original draft, Visualization, Project administration, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Dilxat Muhtar:** Writing – review & editing, Visualization, Methodology, Formal analysis, Data curation, Conceptualization. **Feng Gu:** Writing – review & editing, Visualization. **Yanglangxing He:** Writing – review & editing, Visualization, Validation. **Xueliang Zhang:** Writing – review & editing, Supervision, Resources, Project administration, Funding acquisition. **Pengfeng Xiao:** Writing – review & editing, Supervision, Project administration, Funding acquisition. **Guangjun He:** Writing – review & editing, Resources. **Xiaoxiang Zhu:** Writing – review & editing, Supervision.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This study was supported by the National Natural Science Foundation of China (Grant No. 42071297, 42471410) and the AI and AI for Science Project of Nanjing University (Grant No. 020914380141, D040303). We also would like to thank the High Performance Computing Center of Nanjing University for their help on computation resources.

References

Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F.L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., et al., 2023. Gpt-4 technical report. arXiv preprint arXiv:2303.08774.

Bai, J., Bai, S., Yang, S., Wang, S., Tan, S., Wang, P., Lin, J., Zhou, C., Zhou, J., 2023. Qwen-VL: A versatile Vision-Language model for understanding, localization, text reading, and beyond. arXiv preprint arXiv:2308.12966.

Bai, Z., Wang, P., Xiao, T., He, T., Han, Z., Zhang, Z., Shou, M.Z., 2024. Hallucination of multimodal large language models: A survey. arXiv preprint arXiv:2404.18930.

Bolya, D., Fu, C.-Y., Dai, X., Zhang, P., Feichtenhofer, C., Hoffman, J., 2022. Token merging: Your vit but faster. arXiv preprint arXiv:2210.09461.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J.D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al., 2020. Language models are few-shot learners. Adv. Neural Inf. Process. Syst. 33, 1877–1901.

Chen, L., Li, J., Dong, X., Zhang, P., He, C., Wang, J., Zhao, F., Lin, D., 2023b. Sharegpt4v: Improving large multi-modal models with better captions. arXiv preprint arXiv:2311.12793.

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

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

Cheng, Q., Huang, H., Xu, Y., Zhou, Y., Li, H., Wang, Z., 2022. NWPU-captions dataset and MLCA-net for remote sensing image captioning. IEEE Trans. Geosci. Remote Sens. 60, 1–19.

Christie, G., Fendley, N., Wilson, J., Mukherjee, R., 2018. Functional map of the world. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 6172–6180.

Dai, W., Li, J., Li, D., Tiong, A.M.H., Zhao, J., Wang, W., Li, B., Fung, P., Hoi, S., 2023. InstructBLIP: Towards general-purpose vision-language models with instruction tuning. arXiv:2305.06500.

Dai, D., Yang, W., 2011. Satellite image classification via two-layer sparse coding with biased image representation. IEEE Trans. Geosci. Remote Sens. 8 (1), 173–176.

Ding, J., Xue, N., Xia, G.-S., Bai, X., Yang, W., Yang, M.Y., Belongie, S., Luo, J., Datcu, M., Pelillo, M., et al., 2021. Object detection in aerial images: A large-scale benchmark and challenges. IEEE Trans. Pattern Anal. Mach. Intell. 44 (11), 7778–7796.

Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., Letman, A., Mathur, A., Schelten, A., Yang, A., Fan, A., et al., 2024. The llama 3 herd of models. arXiv preprint arXiv:2407.21783.

Fang, A., Ilharco, G., Wortsman, M., Wan, Y., Shankar, V., Dave, A., Schmidt, L., 2022. Data determines distributional robustness in contrastive language image pre-training (clip). In: International Conference on Machine Learning. PMLR, pp. 6216–6234.

Fei, H., Yao, Y., Zhang, Z., Liu, F., Zhang, A., Chua, T.-S., 2024. From multimodal LLM to human-level AI: Modality, instruction, reasoning, efficiency and beyond. In: Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024): Tutorial Summaries. pp. 1–8.

Gadre, S.Y., Ilharco, G., Fang, A., Hayase, J., Smyrnis, G., Nguyen, T., Marten, R., Wortsman, M., Ghosh, D., Zhang, J., et al., 2024. Datacomp: In search of the next generation of multimodal datasets. Adv. Neural Inf. Process. Syst. 36.

Geva, M., Schuster, R., Berant, J., Levy, O., 2021. Transformer feed-forward layers are key-value memories. arXiv:2012.14913.

Guo, X., Lao, J., Dang, B., Zhang, Y., Yu, L., Ru, L., Zhong, L., Huang, Z., Wu, K., Hu, D., et al., 2024. Skysense: A multi-modal remote sensing foundation model towards universal interpretation for earth observation imagery. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 27672–27683.

Helber, P., Bischke, B., Dengel, A., Borth, D., 2019. Eurosat: A novel dataset and deep learning benchmark for land use and land cover classification. IEEE J. Sel. Top. Appl. Earth Obs. Remote. Sens. 12 (7), 2217–2226.

Hendrycks, D., Burns, C., Basart, S., Zou, A., Mazeika, M., Song, D., Steinhardt, J., 2021. Measuring massive multitask language understanding. In: Proceedings of the International Conference on Learning Representations (ICLR).

Hong, D., Zhang, B., Li, X., Li, Y., Li, C., Yao, J., Yokoya, N., Li, H., Ghamisi, P., Jia, X., et al., 2024. SpectralGPT: Spectral remote sensing foundation model. IEEE Trans. Pattern Anal. Mach. Intell..

Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W., 2021. LoRA: Low-rank adaptation of large language models. arXiv:2106.09685.

Hu, Y., Yuan, J., Wen, C., Lu, X., Li, X., 2023. Rsgpt: A remote sensing vision language model and benchmark. arXiv preprint arXiv:2307.15266.

Irvin, J.A., Liu, E.R., Chen, J.C., Dormoy, I., Kim, J., Khanna, S., Zheng, Z., Ermon, S., 2024. Teochat: A large vision-language assistant for temporal earth observation data. arXiv preprint arXiv:2410.06234.

Jiang, A.Q., Sablayrolles, A., Roux, A., Mensch, A., Savary, B., Bamford, C., Chaplot, D.S., de las Casas, D., Hanna, E.B., Bressand, F., Lengyel, G., Bour, G., Lample, G., Lavaud, L.R., Saulnier, L., Lachaux, M.-A., Stock, P., Subramanian, S., Yang, S., Antoniak, S., Scao, T.L., Gervet, T., Lavril, T., Wang, T., Lacroix, T., Sayed, W.E., 2024. Mixtral of experts. arXiv:2401.04088.

Kuckreja, K., Danish, M.S., Naseer, M., Das, A., Khan, S., 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. pp. 27831–27840.

Li, Y., Du, Y., Zhou, K., Wang, J., Zhao, W.X., Wen, J.-R., 2023b. Evaluating object hallucination in large vision-language models. arXiv preprint arXiv:2305.10355.

Li, C., Gan, Z., Yang, Z., Yang, J., Li, L., Wang, L., Gao, J., et al., 2024c. Multimodal foundation models: From specialists to general-purpose assistants. Found. Trends® Comput. Graph. Vis. 16 (1–2), 1–214.

Li, B., Ge, Y., Ge, Y., Wang, G., Wang, R., Zhang, R., Shan, Y., 2024a. SEED-Bench: Benchmarking multimodal large language models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 13299–13308.

Li, C., Ge, Y., Li, D., Shan, Y., 2023a. Vision-language instruction tuning: A review and analysis. Trans. Mach. Learn. Res..

Li, X., Tu, H., Hui, M., Wang, Z., Zhao, B., Xiao, J., Ren, S., Mei, J., Liu, Q., Zheng, H., et al., 2024d. What if we recaption billions of web images with LLaMA-3?. arXiv preprint arXiv:2406.08478.

Li, X., Wen, C., Hu, Y., Yuan, Z., Zhu, X.X., 2024e. Vision-language models in remote sensing: Current progress and future trends. IEEE Geosci. Remote. Sens. Mag..

Li, B., Zhang, Y., Guo, D., Zhang, R., Li, F., Zhang, H., Zhang, K., Li, Y., Liu, Z., Li, C., 2024b. Llava-OneVision: Easy visual task transfer. arXiv:2408.03326.

Liu, Y., Duan, H., Zhang, Y., Li, B., Zhang, S., Zhao, W., Yuan, Y., Wang, J., He, C., Liu, Z., et al., 2023b. Mmbench: Is your multi-modal model an all-around player?. arXiv preprint arXiv:2307.06281.

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

Liu, H., Li, C., Wu, Q., Lee, Y.J., 2024b. Visual instruction tuning. Adv. Neural Inf. Process. Syst. 36.

Liu, F., Lin, K., Li, L., Wang, J., Yacoob, Y., Wang, L., 2023a. Mitigating hallucination in large multi-modal models via robust instruction tuning. In: The Twelfth International Conference on Learning Representations.

Liu, H., Zhang, X., Xu, H., Shi, Y., Jiang, C., Yan, M., Zhang, J., Huang, F., Yuan, C., Li, B., et al., 2024c. MIBench: Evaluating multimodal large language models over multiple images. arXiv preprint arXiv:2407.15272.

Lobry, S., Marcos, D., Murray, J., Tuia, D., 2020. RSVQA: Visual question answering for remote sensing data. IEEE Trans. Geosci. Remote Sens. 58 (12), 8555–8566.

Lu, X., Wang, B., Zheng, X., Li, X., 2018. Exploring models and data for remote sensing image caption generation. IEEE Trans. Geosci. Remote Sens. 56 (4), 2183–2195.

Luo, J., Pang, Z., Zhang, Y., Wang, T., Wang, L., Dang, B., Lao, J., Wang, J., Chen, J., Tan, Y., et al., 2024. Skysensegpt: A fine-grained instruction tuning dataset and model for remote sensing vision-language understanding. arXiv preprint arXiv:2406.10100.

Ma, L., Liu, Y., Zhang, X., Ye, Y., Yin, G., Johnson, B.A., 2019. Deep learning in remote sensing applications: A meta-analysis and review. ISPRS J. Photogramm. Remote Sens. 152, 166–177.

Mitra, C., Huang, B., Darrell, T., Herzig, R., 2024. Compositional chain-of-thought prompting for large multimodal models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 14420–14431.

Moor, M., Banerjee, O., Abad, Z.S.H., Krumholz, H.M., Leskovec, J., Topol, E.J., Rajpurkar, P., 2023. Foundation models for generalist medical artificial intelligence. Nature 616 (7956), 259–265.

Muhtar, D., Li, Z., Gu, F., Zhang, X., Xiao, P., 2024. Lhrs-bot: Empowering remote sensing with vgi-enhanced large multimodal language model. arXiv preprint arXiv:2402.02544.

Myrzakhan, A., Bsharat, S.M., Shen, Z., 2024. Open-LLM-leaderboard: From multi-choice to open-style questions for LLMs evaluation, benchmark, and arena. arXiv preprint arXiv:2406.07545.

Nguyen, T., Gadre, S.Y., Ilharco, G., Oh, S., Schmidt, L., 2024. Improving multimodal datasets with image captioning. Adv. Neural Inf. Process. Syst. 36.

Nguyen, T., Ilharco, G., Wortsman, M., Oh, S., Schmidt, L., 2022. Quality not quantity: On the interaction between dataset design and robustness of clip. Adv. Neural Inf. Process. Syst. 35, 21455–21469.

OpenAI, 2023a. ChatGPT. https://openai.com/blog/chatgpt/.

OpenAI, 2023b. GPT-4v(ision) system card. URL 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., He, C., 2024a. VHM: Versatile and honest vision language model for remote sensing image analysis. arXiv:2403.20213.

Pang, C., Wu, J., Li, J., Liu, Y., Sun, J., Li, W., Weng, X., Wang, S., Feng, L., Xia, G.-S., et al., 2024b. H2RSVLM: Towards helpful and honest remote sensing large vision language model. arXiv preprint arXiv:2403.20213.

Qian, Z., Chen, M., Sun, Z., Zhang, F., Xu, Q., Guo, J., Xie, Z., Zhang, Z., 2024. Simultaneous extraction of spatial and attributional building information across large-scale urban landscapes from high-resolution satellite imagery. Sustain. Cities Soc. 106, 105393.

Ravuri, S., Lenc, K., Willson, M., Kangin, D., Lam, R., Mirowski, P., Fitzsimons, M., Athanassiadou, M., Kashem, S., Madge, S., et al., 2021. Skilful precipitation nowcasting using deep generative models of radar. Nature 597 (7878), 672–677.

Reichstein, M., Benson, V., Camps-Valls, G., Boran, H., Fearnley, C., Kornhuber, K., Rahaman, N., Schöllkopf, B., Tárraga, J.M., Vinuesa, R., et al., 2024. Early warning of complex climate risk with integrated artificial intelligence.

Reichstein, M., Camps-Valls, G., Stevens, B., Jung, M., Denzler, J., Carvalhais, N., Prabhat, F., 2019. Deep learning and process understanding for data-driven Earth system science. Nature 566 (7743), 195–204.

Ricci, R., Bazi, Y., Melgani, F., 2024. Machine-to-machine visual dialoguing with ChatGPT for enriched textual image description. Remote. Sens. 16 (3), 441.

Robinson, J., Rytting, C.M., Wingate, D., 2021. Leveraging large language models for multiple choice question answering.

Sun, L., Chen, J., Li, Q., Huang, D., 2020. Dramatic uneven urbanization of large cities throughout the world in recent decades. Nat. Commun. 11 (1), 5366.

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

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 J. Photogramm. Remote Sens. 184, 116–130.

Tan, C., Cao, Q., Li, Y., Zhang, J., Yang, X., Zhao, H., Wu, Z., Liu, Z., Yang, H., Wu, N., et al., 2023. On the promises and challenges of multimodal foundation models for geographical, environmental, agricultural, and urban planning applications. arXiv preprint arXiv:2312.17016.

Team, G., Anil, R., Borgeaud, S., Wu, Y., Alayrac, J.-B., Yu, J., Soricut, R., Schalkwyk, J., Dai, A.M., Hauth, A., et al., 2023. Gemini: a family of highly capable multimodal models. arXiv preprint arXiv:2312.11805.

Tong, S., Brown, E., Wu, P., Woo, S., Middepogu, M., Akula, S.C., Yang, J., Yang, S., Iyer, A., Pan, X., Wang, A., Fergus, R., LeCun, Y., Xie, S., 2024. Cambrian-1: A fully open, vision-centric exploration of multimodal LLMs. arXiv:2406.16860.

Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., et al., 2023. Llama 2: Open foundation and fine-tuned chat models. arXiv preprint arXiv:2307.09288.

Wang, Z., Prabha, R., Huang, T., Wu, J., Rajagopal, R., 2024b. Skyscript: A large and semantically diverse vision-language dataset for remote sensing. In: Proceedings of the AAAI Conference on Artificial Intelligence. Vol. 38, pp. 5805–5813.

Wang, H., Zhao, S., Qiang, Z., Qin, B., Liu, T., 2024a. Beyond the answers: Reviewing the rationality of multiple choice question answering for the evaluation of large language models. arXiv preprint arXiv:2402.01349.

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., Le, Q.V., Zhou, D., et al., 2022. Chain-of-thought prompting elicits reasoning in large language models. Adv. Neural Inf. Process. Syst. 35, 24824–24837.

Wu, S., Chen, B., Webster, C., Xu, B., Gong, P., 2023. Improved human greenspace exposure equality during 21st century urbanization. Nat. Commun. 14 (1), 6460.

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

Xiong, Z., Wang, Y., Zhang, F., Stewart, A.J., Hanna, J., Borth, D., Papoutsis, I., Saux, B.L., Camps-Valls, G., Zhu, X.X., 2024. Neural plasticity-inspired foundation model for observing the earth crossing modalities. arXiv preprint arXiv:2403.15356.

Xu, Z., Feng, C., Shao, R., Ashby, T., Shen, Y., Jin, D., Cheng, Y., Wang, Q., Huang, L., 2024b. Vision-flan: Scaling human-labeled tasks in visual instruction tuning. arXiv preprint arXiv:2402.11690.

Xu, Q., Shi, Y., Bamber, J.L., Ouyang, C., Zhu, X.X., 2024a. Large-scale flood modeling and forecasting with FloodCast. Water Res. 122162.

Xu, H., Xie, S., Tan, X.E., Huang, P.-Y., Howes, R., Sharma, V., Li, S.-W., Ghosh, G., Zettlemoyer, L., Feichtenhofer, C., 2023. Demystifying clip data. arXiv preprint arXiv:2309.16671.

Yang, Y., Newsam, S., 2010. 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. pp. 270–279.

Ye, Q., Xu, H., Ye, J., Yan, M., Hu, A., Liu, H., Qian, Q., Zhang, J., Huang, F., 2024. Mplug-owl2: Revolutionizing multi-modal large language model with modality collaboration. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 13040–13051.

Yin, S., Fu, C., Zhao, S., Li, K., Sun, X., Xu, T., Chen, E., 2023. A survey on multimodal large language models. arXiv preprint arXiv:2306.13549.

Yuan, Z., Xiong, Z., Mou, L., Zhu, X.X., 2024. Chatearthnet: A global-scale image-text dataset empowering vision-language geo-foundation models. Earth Syst. Sci. Data Discuss. 2024, 1–24.

Yuan, Z., Zhang, W., Fu, K., Li, X., Deng, C., Wang, H., Sun, X., 2022. Exploring a fine-grained multiscale method for cross-modal remote sensing image retrieval. IEEE Trans. Geosci. Remote Sens. 60, 1–19.

Zhai, X., Mustafa, B., Kolesnikov, A., Beyer, L., 2023. Sigmoid loss for language image pre-training. arXiv:2303.15343.

Zhan, Y., Xiong, Z., Yuan, Y., 2023. Rsvg: Exploring data and models for visual grounding on remote sensing data. IEEE Trans. Geosci. Remote Sens. 61, 1–13.

Zhan, Y., Xiong, Z., Yuan, Y., 2024. Skyeyegpt: Unifying remote sensing vision-language tasks via instruction tuning with large language model. arXiv preprint arXiv:2401.09712.

Zhang, W., Cai, M., Zhang, T., Zhuang, Y., Mao, X., 2024b. Earthgpt: A universal multi-modal large language model for multi-sensor image comprehension in remote sensing domain. IEEE Trans. Geosci. Remote Sens..

Zhang, S., Dong, L., Li, X., Zhang, S., Sun, X., Wang, S., Li, J., Hu, R., Zhang, T., Wu, F., et al., 2023b. Instruction tuning for large language models: A survey. arXiv preprint arXiv:2308.10792.

Zhang, C., Wang, S., 2024. Good at captioning, bad at counting: Benchmarking gpt-4v on earth observation data. arXiv preprint arXiv:2401.17600.

Zhang, P., Wang, X.D.B., Cao, Y., Xu, C., Ouyang, L., Zhao, Z., Ding, S., Zhang, S., Duan, H., Yan, H., et al., 2023a. Internlm-xcomposer: A vision-language large model for advanced text-image comprehension and composition. arXiv preprint arXiv:2309.15112.

Zhang, Z., Xu, L., Jiang, Z., Hao, H., Wang, R., 2024c. Multiple-choice questions are efficient and robust LLM evaluators. arXiv preprint arXiv:2405.11966.

Zhang, L., Yang, L., Zohner, C.M., Crowther, T.W., Li, M., Shen, F., Guo, M., Qin, J., Yao, L., Zhou, C., 2022. Direct and indirect impacts of urbanization on vegetation growth across the world's cities. Sci. Adv. 8 (27), eabo0095.

Zhang, L., Zhang, L., 2022. Artificial intelligence for remote sensing data analysis: A review of challenges and opportunities. IEEE Geosci. Remote. Sens. Mag. 10 (2), 270–294.

Zhang, B., Zhang, P., Dong, X., Zang, Y., Wang, J., 2024a. Long-CLIP: Unlocking the long-text capability of CLIP. arXiv preprint arXiv:2403.15378.

Zhang, Z., Zhao, T., Guo, Y., Yin, J., 2023. Rs5m: A large scale vision-language dataset for remote sensing vision-language foundation model. arXiv preprint arXiv:2306.11300.

Zhao, T., Wang, S., Ouyang, C., Chen, M., Liu, C., Zhang, J., Yu, L., Wang, F., Xie, Y., Li, J., et al., 2024. Artificial intelligence for geoscience: Progress, challenges and perspectives. Innov..

Zhao, W.X., Zhou, K., Li, J., Tang, T., Wang, X., Hou, Y., Min, Y., Zhang, B., Zhang, J., Dong, Z., et al., 2023. A survey of large language models. arXiv preprint arXiv:2303.18223.

Zheng, L., Chiang, W.-L., Sheng, Y., Zhuang, S., Wu, Z., Zhuang, Y., Lin, Z., Li, Z., Li, D., Xing, E., et al., 2024. Judging llm-as-a-judge with mt-bench and chatbot arena. Adv. Neural Inf. Process. Syst. 36.

Zhou, Y., Feng, L., Ke, Y., Jiang, X., Yan, J., Yang, X., Zhang, W., 2024. Towards vision-language geo-foundation model: A survey. arXiv preprint arXiv:2406.09385.

Zhu, B., Lui, N., Irvin, J., Le, J., Tadwalkar, S., Wang, C., Ouyang, Z., Liu, F.Y., Ng, A.Y., Jackson, R.B., 2022a. Meter-ml: A multi-sensor earth observation benchmark for automated methane source mapping. arXiv preprint arXiv:2207.11166.

Zhu, Z., Qiu, S., Ye, S., 2022b. Remote sensing of land change: A multifaceted perspective. Remote Sens. Environ. 282, 113266.

Zhu, X.X., Tuia, D., Mou, L., Xia, G.-S., Zhang, L., Xu, F., Fraundorfer, F., 2017. Deep learning in remote sensing: A comprehensive review and list of resources. IEEE Geosci. Remote. Sens. Mag. 5 (4), 8–36.

Zhu, X.X., Xiong, Z., Wang, Y., Stewart, A.J., Heidler, K., Wang, Y., Yuan, Z., Dujardin, T., Xu, Q., Shi, Y., 2024. On the foundations of earth and climate foundation models. arXiv preprint arXiv:2405.04285.

Zhu, Q., Zhong, Y., Zhao, B., Xia, G.-S., Zhang, L., 2016. Bag-of-visual-words scene classifier with local and global features for high spatial resolution remote sensing imagery. IEEE Geosci. Remote. Sens. Lett. 13 (6), 747–751.