遥感多模态大语言模型：架构、关键技术和未来展望

许文嘉*[①] 于睿卿[①] 薛铭浩[①] 张源奔[②] 魏智威[③] 张 柘[②④⑤⑥] 彭木根[①]

[①](北京邮电大学网络与交换技术全国重点实验室 北京 100081)

[②](中国科学院空天信息创新研究院 北京 100091)

[③](湖南师范大学地理科学学院 长沙 410081)

[④](苏州空天信息研究院 苏州 215123)

[⑤](微波成像全国重点实验室 北京 100190)

[⑥](中国科学院大学电子电气与通信工程学院 北京 100190)

摘要：近年来，人工智能技术和遥感领域的结合已成为领域发展的前沿热点，多模态大语言模型(MLLM)的快速发展为遥感智能解译带来新的机遇和挑战。遥感多模态大语言模型通过构建大语言模型与视觉模型之间的桥接机制并采用联合训练方式，深度融合遥感领域的视觉特征与语义信息，有效推动遥感智能解译由浅层语义匹配向高层的世界知识理解跃迁。该文系统性回顾了多模态大语言模型在遥感领域的相关研究成果，以期为新的研究方向提供依据。具体而言，该文首先明确了遥感多模态大语言模型(RS-MLLM)的概念定义，并梳理了遥感多模态大语言模型的发展脉络。随后，详细阐述了遥感多模态大语言模型的模型架构、训练方法、适用任务及其对应的基准数据集，并介绍了遥感智能体。最后，探讨了遥感多模态大语言模型的研究现状和未来发展方向。

关键词：大语言模型；多模态大语言模型；遥感多模态大语言模型；视觉语言模型；遥感智能体

中图分类号：TN957.51; TP753 **文献标识码**：A **文章编号**：2095-283X(2025)x-0001-24

DOI: 10.12000/JR25088 **CSTR**: 32380.14.JR25088

引用格式：许文嘉, 于睿卿, 薛铭浩, 等. 遥感多模态大语言模型：架构、关键技术和未来展望[J]. 雷达学报(中英文), 待出版. doi: 10.12000/JR25088.

Reference format: XU Wenjia, YU Ruiqing, XUE Minghao, *et al*. A survey on remote sensing multimodal largelanguage models: Framework, core technologies, and future perspectives[J]. *Journal of Radars*, in press. doi: 10.12000/JR25088.

A Survey on Remote Sensing Multimodal Large Language Models: Framework, Core Technologies, and Future Perspectives

XU Wenjia*[①] YU Ruiqing[①] XUE Minghao[①] ZHANG Yuanben[②] WEI Zhiwei[③]
ZHANG Zhe[②④⑤⑥] PENG Mugen[①]

[①](*State Key Laboratory of Networking and Switching Technology*, *Beijing University of Posts and Telecommunications*, *Beijing* 100081, *China*)

[②](*Aerospace Information Research Institute*, *Chinese Academy of Sciences*, *Beijing* 100091, *China*)

[③](*School of Geographical Sciences*, *Hunan Normal University*, *Changsha* 410081, *China*)

收稿日期：2025-05-12；改回日期：2025-07-22；网络出版：2025-xx-xx

*通信作者： 许文嘉 xuwenjia@bupt.edu.cn *Corresponding Author: XU Wenjia, xuwenjia@bupt.edu.cn

基金项目：国家自然科学基金(62301063)，目标认知与应用技术重点实验室开放基金(2023-CXPT-LC-005)，微波成像技术国家重点实验室开放基金(70323006)

Foundation Items: The National Natural Science Fundation of China (62301063), The Key Laboratory of Target Cognition and Application Technology (2023-CXPT-LC-005), The National Key Laboratory of Microwave Imaging Technology (70323006)

责任主编：张帆 Corresponding Editor: ZHANG Fan

④(*Suzhou Aerospace Information Research Institute*, *Suzhou* 215123, *China*)
⑤(*Science and Technology on Microwave Imaging Laboratory*, *Beijing* 100190, *China*)
⑥(*Electrical and Communication Engineering*, *University of Chinese Academy of Sciences*, *Beijing* 100190, *China*)

Abstract: In recent years, the rapid development of Multimodal Large Language Models (MLLMs) and their applications in remote sensing have garnered significant attention. Remote sensing MLLMs achieve deep integration of visual features and semantic information through the design of bridging mechanisms between large language models and vision models, combined with joint training strategies. This integration facilitates a paradigm shift in intelligent remote sensing interpretation—from shallow semantic matching to higher-level understanding based on world knowledge. In this study, we systematically review the research progress in the applications of MLLMs in remote sensing, specifically examining the development of Remote Sensing MLLMs (RS-MLLMs), which provides a foundation for future research directions. Initially, we discuss the concept of RS-MLLMs and review their development in chronological order. Subsequently, we provide a detailed analysis and statistical summary of the proposed architectures, training methods, applications, and corresponding benchmark datasets, along with an introduction to remote sensing agents. Finally, we summarize the research status of RS-MLLMs and discuss future research directions.

Key words: Large Language Model (LLM); Multimodal Large Language Model (MLLM); Remote Sensing Multimodal Large Language Model (RS-MLLM); Vision Language Model (VLM); Remote sensing agent

1 引言

遥感技术通过传感器远距离收集地表信息，能够高效提供大范围地表实时监测能力，已广泛应用于环境监测[1]、灾害评估[2]、农业生产[3]等众多领域。随着机器学习和深度学习的发展，大量遥感图像解译应用算法应运而生[4–6]。但是，上述传统算法通常针对单一任务或单一模态数据进行优化，难以应对遥感领域中任务目标和数据模态的多样性。随着遥感大数据时代的到来，面向爆炸式增长的遥感数据和多样化遥感应用，传统算法在泛化能力和多模态数据处理能力方面的局限性更加显现[7]。因此，提升模型的泛化能力并实现多模态信息的高效融合，已成为当前遥感领域亟待解决的关键问题。

近年来，大语言模型(Large Language Model, LLM)[8]和多模态大语言模型(Multimodal Large Language Model, MLLM)[9]的发展为遥感领域带来新的解决方案。LLM通过大规模预训练获得了优异的语义理解和泛化能力，能够处理和理解用户多样化的需求；MLLM则在保留LLM语义理解和泛化能力的同时，增加了处理文本、图像、视频等多模态数据的能力。因此，若将遥感数据也纳入MLLM的应用范畴，就能有效借助MLLM的跨模态信息融合与理解能力完成多样化的多模态遥感任务，推动遥感技术的智能化进程。在这一大背景下，遥感多模态大语言模型(Remote Sensing Multimodal Large Language Model, RS-MLLM)应运而生。RS-MLLM是指针对遥感领域优化的MLLM，其能够接收并处理包括遥感图像(如光学图像、合成孔径雷达影像)、卫星视频以及相关文本描述等在内的多模态输入，并在LLM的支持下完成相应的遥感领域应用任务。该模型融合了LLM的语义理解能力与多模态技术的信息处理能力，已在遥感图像描述(Remote Sensing Image Captioning, RSIC)、遥感视觉问答(Remote Sensing Visual Question Answering, RSVQA)、遥感视觉定位(Remote Sensing Visual Grounding, RSVG)、遥感场景分类(Remote Sensing Scene Classification, RSSC)等多种任务中表现出强大的泛化能力[10–13]，其发展为提升遥感任务的准确性、效率以及处理复杂遥感问题提供了强有力的技术支持。

传统遥感解译模型主要依赖视觉模型在影像与语义标签之间建立映射关系。与之不同，RS-MLLM通过构建大语言模型与视觉模型之间的桥接机制，深度融合遥感领域的视觉信息与世界知识，有效推动遥感智能解译由浅层语义匹配向深层世界知识理解的跃迁，进而引发遥感解译范式的新一轮变革。

因此，为促进RS-MLLM的发展，本文对RS-MLLM进行系统性回顾，旨在为研究人员提供有价值的参考。具体而言，本文明确了LLM, MLLM及RS-MLLM的相关概念，并梳理了它们的发展脉络。随后详细分析了现有RS-MLLM的架构、训练方法、适用任务及其对应的基准数据集。最后，对RS-MLLM的发展方向进行了展望。

2 相关概念与发展历程

本节首先介绍大语言模型、多模态大语言模型

及遥感多模态大语言模型的相关概念，并按照时间顺序梳理RS-MLLM的发展历程。

2.1 相关概念

2.1.1 大语言模型

语言模型的研究由来已久，初期研究主要集中于基于规则的自然语言处理方法[14,15]。后续，随着计算能力的提高和数据量的增加，基于统计的语言模型逐渐成为主流。进入21世纪，随着计算机技术的迅猛发展，深度学习有关的技术则被广泛应用于语言模型[16]。但是，上述语言模型通常规模较小，且多只能处理单一语言任务，如句法分析、情感分类等。2018年，BERT (Bidirectional Encoder Representations from Transformers)[17]模型的发布则是语言模型发展的一个重要里程碑，BERT通过双向Transformer架构的创新，使得模型能够高效地捕捉长距离依赖关系和上下文信息，显著提升了模型自然语言理解任务的效果。随之而来的基于预训练和微调的GPT (Generative Pre-trained Transformer)[18–21]系列模型，进一步推动了自然语言生成任务的进展，使得语言模型进入到大模型时代，并催生了一系列相关的LLMs，如GPT-3[20], LLaMA[22], QWEN[23], DeepSeekV3[24]等。

因此，本文提到的LLM特指通过深度学习方法(尤其是基于Transformer[25]架构的神经网络模型)来理解和生成自然语言文本的具有大规模参数的人工智能系统。其核心思想是通过大规模的语料库进行预训练，使模型能够捕捉语言中的语法、语义和上下文关系。与传统的语言模型不同，LLM不仅能够处理单一的文本输入，还能够在生成文本时保持上下文的连贯性，并进行长文本的理解和生成等。

2.1.2 多模态大语言模型

MLLM是在大模型时代发展的用于处理多模态数据的深度学习模型，其使用LLM作为核心，能够处理多种模态数据[9]，典型代表为OpenAI公司推出的GPT4o。尽管该模型能够处理文本、图像、视频、音频等多种输入，但其核心任务或输出通常是围绕语言处理展开的，主要包括将其他模态(如图像、视频、音频等)转化为语言形式(例如生成文本、回答问题或生成描述等)[26,27]，或利用语言来引导其他模态(例如从文本描述生成图像)[28]。相比于LLM仅限于对文本信息进行处理，MLLM在保留了自然语言这一灵活交互形式的同时，通过融合不同模态的信息，显著提升了模型的理解和生成能力。

典型的MLLM结构如图1所示，包括多模态编码器(3.1节)、多模态特征投影器(3.2节)、预训练大语言模型(3.3节)以及多模态解码器。其中，多模态编码器负责接收并编码非文本信息，如图像、音频和视频等。多模态特征投影器则负责将多模态编码器提取的特征通过投影、查询和融合等方式映射到语义空间，以便与文本嵌入合并后输入到预训练大语言模型中。最终，预训练的大语言模型对多模态信息进行融合、理解和回复。多模态解码器则基于文本输出生成其他模态信息，通常为可选项。

2.1.3 遥感多模态大语言模型

遥感多模态大模型是MLLM面向遥感领域应用的延伸，RS-MLLM亦遵循典型MLLM架构，并专门针对遥感领域进行定制和优化[12,13,29]。RS-MLLM具备面向遥感多模态数据的处理与交互能力，支持文本、图像(包括光学图像、合成孔径雷达影像等多源遥感图像)、卫星视频等多模态输入，并在LLM的支持下完成典型的遥感领域应用任务，如遥感图像描述、遥感视觉问答、遥感视觉定位等[29–31]。

随着人工智能技术的飞速发展，遥感领域大模型有关的研究亦进展迅猛，出现了与RS-MLLM相近的多个概念，如多模态遥感基础大模型、遥感视觉-语言模型等。为了更好厘清本文RS-MLLM

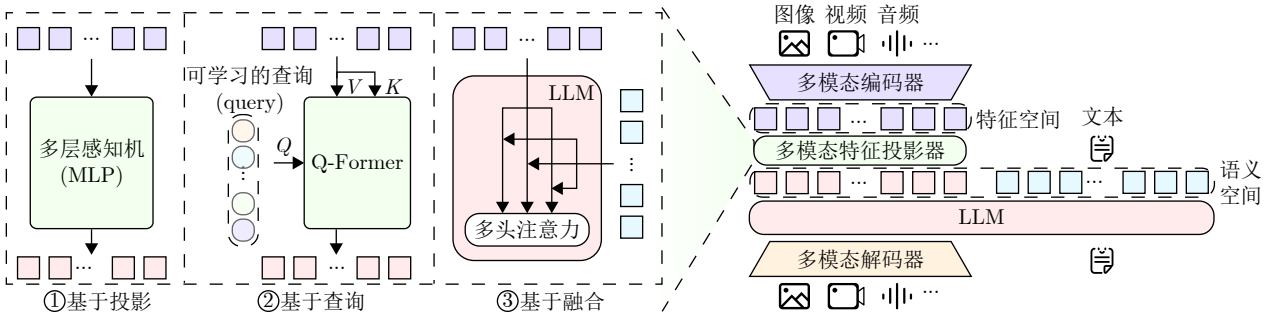

图 1 典型的多模态大语言模型架构

Fig. 1 The typical MLLM architecture

的概念和边界范围，在此，对RS-MLLM与多模态遥感基础大模型以及遥感视觉-语言模型等概念进行区分。

(1) 遥感视觉基础大模型[32]：其核心目标是通过大量未标注的遥感数据进行预训练以构建一个通用的基础模型，用以学习遥感数据的统一特征表示。遥感基础大模型的重点在于自监督学习，其通过从SAR(Synthetic Aperture Radar)、可见光等多种遥感数据中学习和提取通用的特征表达，为后续的具体任务提供基础特征支持，例如SpectralFormer[33]通过改进的Transformer架构，专注于高光谱图像分类，采用分组光谱嵌入(GSE)和跨层自适应融合(CAF)模块，优化了光谱序列的捕捉和信息传递，从而提升了分类精度；而SpectralGPT[34]也是针对光谱遥感图像进行预训练，显著提升了模型在光谱遥感场景分类、语义分割和变化检测等多种下游任务中的表现。这些基础模型主要是针对特定的图像进行特征提取，为下游任务提供支持。而多模态遥感基础大模型则是指针对不同模态的数据分别建立相应的遥感基础大模型，以确保模型能够有效处理各种模态的数据特性[35,36]。

(2) 遥感多模态视觉-语言模型[37]：其主要关注图像与文本这两种模态的交互，通常采用对比学习或数据融合方法，将视觉特征与语言特征进行对齐。通过对遥感影像及其对应的语言描述进行比对或融合，该模型能够在视觉与语言之间建立紧密的映射关系。然而，由于其主要面向图像与文本两大模态，对其他模态数据的处理能力相对有限，因此主要应用于图像-文本相关的任务。

(3) 遥感多模态大语言模型[10]：其在架构设计上以LLM为核心，通过结合多模态编码器和多模态特征投影器，能实现遥感数据的多模态融合处理。该模型不仅强调多模态遥感数据作为输入，还在输出上强调围绕LLM展开应用，包括将其他模态数据(如遥感影像、遥感视频)转化为语言形式(例如遥感图像描述任务和变化检测任务)，或者利用语言来引导其他模态生成(例如图像分割任务)等。

2.2 遥感多模态大语言模型发展历程

RS-MLLM是大模型时代遥感领域的重大进展，大模型的崛起则源于自然语言处理领域的深刻变革。随着计算能力的提升、海量数据的积累和深度学习算法的不断优化，研究者逐步突破了传统技术的局限，使得基于大规模预训练的语言模型成为推动人工智能发展的核心技术之一。OpenAI推出的GPT系列模型，作为这一领域的重要代表，标志着大语言模型技术的重大突破。特别是GPT-3[20]，其通过大规模预训练展示了强大的自然语言生成与理解能力，为大语言模型技术的进一步发展奠定了基础。随后的GPT-4[21]在参数规模和性能上实现了显著突破，提升了语言理解和生成的准确性与流畅度。此外，Meta推出的LLaMA[22]和Vicuna[38]、Anthropic公司推出的Claude等模型也在各自的优势领域表现突出，推动了LLM多样化的发展。而作为近年来涌现的新兴大模型之一，DeepSeek[39]凭借其强大的推理能力、高效的训练架构和在多模态任务上的探索，也为大语言模型的发展提供了新的方向和可能性。

随着大语言模型能力的不断增强，研究人员对于MLLM的研究热情也日益高涨，旨在通过将多模态信息输入LLM，提升人工智能在复杂任务中的表现。例如，Flamingo[40]通过视觉和语言的融合，能够在少量示例下完成视觉问答任务；BLIP-2[41]利用引导式学习，在图像描述生成和视觉问答等任务中取得了显著成效；LLaVA[26]通过将视觉信息与语言模型结合，在图像问答、图像描述生成等任务中表现出色；Video-LLaVA[27]则在LLaVA的基础上进一步拓展，能够处理视频数据并完成视频问答和视频内容生成等任务。上述模型进一步推动了多模态技术的发展，并逐步扩展至医学、生物等多个领域。

RS-MLLM则是上述研究趋势向遥感领域的扩展，通过在遥感数据处理中引入MLLM，研究人员能够实现对多模态遥感数据的综合分析与理解，并迅速引起了广泛关注。但是遥感领域与自然领域的信息分布有较大的差距，为了缓解领域漂移的难题，专门针对遥感领域进行定制和优化的RS-MLLM被陆续提出，相关的RS-MLLM发展脉络梳理见图2。其中，RSGPT[10]是第1个支持图像描述生成和视觉问答任务的RS-MLLM。GeoChat[42]则是首个支持视觉定位任务的对话式RS-MLLM。SkyEyeGPT[31]则是首个实现遥感视频描述任务的RS-MLLM。VHM[43]进一步引入诚信度概念的模型，能够拒绝回答不确定的问题。SkySenseGPT[44]则引入了图谱的概念，增强了RS-MLLM在感知和推理遥感目标之间关系以及复杂理解任务中的能力。RS-CapRet[45]则是首个实现遥感图像检索任务的遥感多模态大语言模型。IFship[46]和Popeye[47]则专注于船只检测任务，ChangeChat[48]则作为首个面向遥感图像变化描述任务的模型，填补了该领域的研究空白。随后提出的CDChat[49]和TEOChat[50]进一步从深度与广度上丰富了这一领域的研究方法与应用场景。UniRS[13]则作

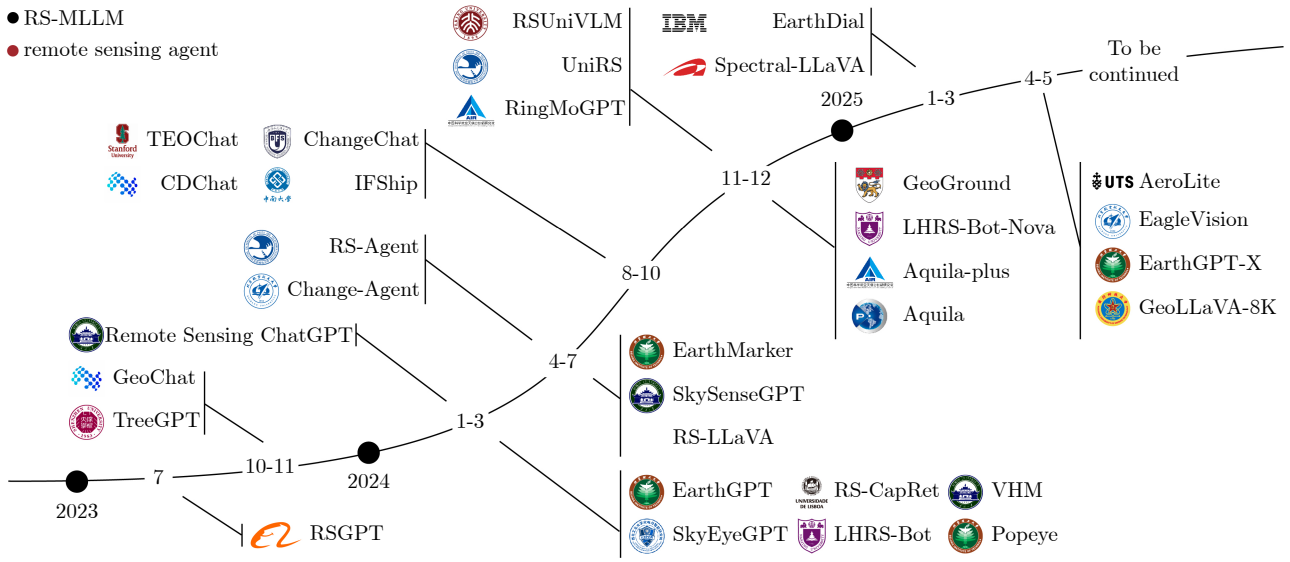

图 2 代表性遥感多模态大语言模型和遥感智能体的发展时间线

Fig. 2 Timeline of representative RS-MLLMs and remote sensing agents

为首个专为解决多时相遥感任务而设计的RS-MLLM，融合了遥感中3种关键的时相视觉输入类型(即单张图像、双时相图像对和视频)，拓展了RS-MLLM在时序遥感分析任务中的能力。

3　遥感多模态大语言模型架构

RS-MLLM亦遵循典型MLLM的架构，如图1所示。目前的RS-MLLM主要由多模态编码器、多模态投影器和预训练大语言模型3个核心模块组成。因此，本节将重点对RS-MLLM中使用的多模态编码器(3.1节)、多模态投影器(3.2节)和预训练大语言模型(3.3节)进行分析和介绍。表1总结了不同RS-MLLM模型的具体结构，这些模型在多模态编码器、特征投影器和预训练大语言模型的选择上呈现出较高的相似性，具体论述如下。

3.1　多模态编码器

多模态编码器主要负责处理多模态遥感数据(如图像、视频、音频等)的输入，并将不同模态的信息映射至统一的表示空间，以实现后续模块跨模态任务的高效处理。现有的RS-MLLM多采用基于对比学习预训练的CLIP模型[51]作为多模态编码器。CLIP利用独立的视觉和文本编码器提取图像和文本特征，将它们映射到共享的高维空间。通过对比学习，CLIP确保相关图像和文本被映射到相似的向量位置，最大化相似模态间的相似度，最小化不相关模态间的距离，从而在遥感任务中提供强大的跨模态理解能力，支持图像检索、文本生成等任务。随着应用需求的不断发展，现有的CLIP模型有多个变体，它们在视觉编码器架构上进行了不同程度的改进，旨在提升模型在不同场景下的表现。具体来说，CLIP-ViT[51], CLIP-ConvNeXt[52]和EVA-CLIP[53]等变体采用了不同的视觉编码器架构，各自具有不同的特征提取能力与优势，因而在不同应用场景中展现出更好的性能。

CLIP-ViT使用ViT(Vision Transformer)[54]作为视觉特征编码器。ViT将图像切分为若干小块并输入到Transformer模型中进行处理，通过自注意力机制捕捉图像中的全局信息，适合处理那些需要全局建模的任务，尤其是在遥感图像分析中，能够有效地捕捉长距离依赖信息，因此在处理大规模和复杂数据集时表现尤为突出。CLIP-ViT在遥感多模态大语言模型中应用最为广泛，CDChat[49], ChangeChat[48]和GeoChat[42]等模型均采用这一多模态编码器。

CLIP-ConvNeXt模型则采用了改进版的ConvNeXt架构，ConvNeXt主要通过优化卷积层设计、融合Transformer架构中的设计理念、增加网络深度以及提高计算效率来提升传统CNN的性能。这使得CLIP-ConvNeXt在处理细节、局部特征和计算效率上表现更好，特别适用于需要精细化局部特征提取的任务。在Aquila[55]和Aquila-plus[56]等模型中使用。

EVA-CLIP则结合了ViT的全局建模能力和针对计算效率优化的EVA(Efficient Vision Transformer)架构。EVA主要通过优化自注意力机制、高效的Transformer结构、减少冗余计算、参数优化和内存优化等方面的改进，使得ViT在计算效率、推理速度和资源利用率上得到了显著提升。这使得EVA-CLIP适合资源受限的环境或大规模数据

表 1 遥感多模态大语言模型具体结构

Tab. 1 The specific structure of RS-MLLM

RS-MLLM	多模态编码器	多模态特征投影器	预训练大语言模型	训练硬件
AeroLite[57]	CLIP ViT-L/14	a two-layer MLP	LLaMA3.2-3B	4090
Aquila[55]	Aquila-CLIP ConvNext (A-CCN)	SFI	(MDA)-LLM (基于LLaMA3)	4*A800
Aquila-plus[56]	CLIP ConvNeXt-L	Mask Spatial Feature Extractor	Vicuna	-
CDChat[49]	CLIP ViT-L/14	a two-layer MLP(GELU)	Vicuna-v1.5-7B	3*A100
ChangeChat[48]	CLIP ViT/14	a two-layer MLP	Vicuna-v1.5	L20(48 GB)
EagleVision[58]	Baseline Detector	Attribute Disentangle	InternLM2.5-7B-Chat等	8*A100
EarthDial[59]	Adaptive High Resolution + Data Fusion + InternViT-300M	a simple MLP	Phi-3-mini	8*A100(80 GB)
EarthGPT[12]	DINOv2 ViT-L/14 + CLIP ConvNeXt-L	a linear layer	LLaMA2-13B	16*A100
EarthGPT-X[60]	DINOv2-ViT L/14 + CLIP ConvNeXt-L + Hybrid Signals Mutual Understanding	Vision-to-Language Modality-align Projection	LLaMA2-13B	8*A100(80 GB)
EarthMarker[61]	DINOv2 ViT-L/14 + CLIP ConvNeXt-L	a linear layer	LLaMA2-13B	8*A100(80 GB)
GeoChat[42]	CLIP ViT-L/14	a two-layer MLP(GELU)	Vicuna-v1.5-7B	-
GeoGround[29]	CLIP ViT	a two-layer MLP	Vicuna-v1.5	8*V100(32 GB)
GeoLLaVA-8K[62]	CLIP ViT-L/14 + a two-step tokens compression module	a linear layer	Vicuna-v1.5-7B	-
IFShip[46]	CLIP ViT-L/14	a four-layer MLP(GELU)	Vicuna-13B	-
LHRS-Bot[11]	CLIP ViT-L/14	Vision Perceiver	LLaMA2-7B	8*V100(32 GB)
LHRS-Bot-Nova[63]	SigLIP-L/14	Vision Perceiver	LLaMA3-8B	8*H100
Popeye[47]	DINOv2 ViT-L/14 + CLIP ConvNeXt-L	Alignment Projection	LLaMA-7B	-
RingMoGPT[30]	EVA-CLIP ViT-g/14	a Q-Former + a linear layer	Vicuna-13B	8*A100(80 GB)
RS-CapRet[45]	CLIP ViT-L/14	three linear layers	LLaMA2-7B	-
RSGPT[10]	EVA-G	a Q-Former + a linear layer	Vicuna-7B Vicuna-13B	8*A100
RS-LLaVA[64]	CLIP ViT-L	a two-layer MLP(GELU)	Vicuna-v1.5-7B Vicuna-v1.5-13B	2*A6000(48 GB)
RSUniVLM[65]	SigLIP-400M	a two-layer MLP	QWen2-0.5B	4*A40(40 GB)
SkyEyeGPT[31]	EVA-CLIP	a linear layer	LLaMA2	4*3090
SkySenseGPT[44]	CLIP ViT-L/14	a two-layer MLP	Vicuna-v1.5	4*A100(40 GB)
Spectral-LLaVA[66]	SpectralGPT[34] (encoder only)	a linear layer	LLaMA3	-
TEOChat[50]	CLIP ViT-L/14	a two-layer MLP	LLaMA2	A4000(16 GB)
UniRS[13]	SigLIP + a Change Extraction Module	a downsampling module + a MLP	Sheared-LLaMA (3B)	4*4090(24 GB)
VHM[43]	CLIP ViT-L/14	a two-layer MLP	Vicuna-v1.5-7B	16*A100(80 GB)

注：斜体表示该模型在论文中并未给出正式名称或缩写。

集处理，在RingMoGPT[30]和SkyEyeGPT[31]模型中作为多模态编码器使用。

此外，考虑到不同模型在网络架构上的差异，有时多个多模态编码器会被同时使用，为不同模型提供互补的视觉语义。例如EarthGPT[12], EarthMarker[61]和Popeye[47]等，同时使用了基于ViT的DINO[67]模型和基于ConvNeXt的CLIP模型作为多模态编码器，这有助于进一步提升视觉语义的表达能力和跨模态任务的性能。除上述多模态编码器外，还有一些其他的模型也被用于多模态编码，例如SigLIP模型[68]则被应用在LHRS-Bot-Nova[63]中。

除上述模型外，近年来MoE编码器也被广泛应用于多模态任务，尤其是在需要处理复杂和多样化数据集时，通过激活不同的专家模型来提高计算效率并保持较高的性能。这类模型在图像理解、生成和处理任务中展现出了强大的能力。例如，CLIP-MoE[69]通过将多个专家模型结合在一起，能够在视觉和文本的多模态任务中提供更精确的特征表达，同时避免了传统模型在计算资源和训练时间上的高昂成本。未来，基于MoE架构的编码器有望作为高效的多模态特征编码器被广泛应用于遥感多模态大语言模型中，推动该领域的发展。

3.2 多模态特征投影器

多模态特征投影器是为了建立多模态编码器获取的特征空间与后续预训练大语言模型语义信息之间的联系，即使用可学习的多模态特征投影器将多模态特征信息全部映射至语义空间，便于后续预训练大语言模型进行处理。基于特征映射方式的不同，多模态特征投影器主要包括3种方式：基于投影、基于查询和基于融合的方式，其架构如图1左侧所示。

(1) 基于投影的多模态特征投影器：其主要是通过多层感知机(Multilayer Perceptron, MLP)将多模态信息映射到语义空间。以输入图像文本对为例，图像$I \in R^{H\times W\times 3}$通过视觉编码器$E$提取出图像特征向量$F_v = E(I)$，随后图像特征通过一个可学习的多层感知机映射到语言空间，即

$$F_v' = \mathrm{MLP}(F_v) \tag{1}$$

投影后的视觉特征F_v'被作为额外上下文，与任务指令和文本输入拼接后共同送入大语言模型中，以生成输出。当前大多数RS-MLLM，如RS-LLaVA[64], SkySenseGPT[44]和VHM[43]均采用类似的设计策略。具体来说，MLP通过层叠的非线性变换来实现特征空间与语义空间的映射，使得模型能够更好地理解多模态数据。在此基础上，一些模型，如CDChat[49], GeoChat[42], IFShip[46]和RS-LLaVA[64]则通过在MLP中引入GeLU激活函数，进一步优化了多模态特征投影器的性能表现。

(2) 基于查询的多模态特征投影器：主要是通过一组可学习的查询来提取信息，仍以输入图像文本对为例，为了从图像特征向量中提取与任务语义相关的视觉表示，引入一组可学习的查询向量Q，通过交叉注意力机制在Q-Former中与图像特征F_v进行动态交互，得到结构化、具有任务感知能力的视觉表示：

$$F_v' = Q\text{-Former}(F_v, Q) \tag{2}$$

得到的图像特征F_v'同样与任务指令和文本输入拼接后共同送入大语言模型中，以生成输出。这种方法首先在BLIP-2中使用，部分RS-MLLM也采用了这种方式，如RingMoGPT[30]和RSGPT[10]等。在这一框架下，查询作为一种可学习的表示，能够引导模型提取出关键的视觉语义信息以实现特征空间和语义的映射。此外，Aquila[55]中的SFI模块以及LHRS-Bot[11]和LHRS-Bot-Nova[63]中的Vision Perceiver模块也采用了这种方式。

(3) 基于融合的多模态特征投影器：其主要是通过在大语言模型的冻结Transformer层之间插入额外的交叉注意力层，从而引入外部视觉线索以增强语言特征的表达。仍以输入图像文本对为例，具体是将处理后的视觉表示与文本信息拼接后输入至一个融合式 Transformer 中，在其中进行跨模态的交叉注意力建模，即同时建模图文token间的相互关系，表示为

$$h = \mathrm{Transformer}(F_v, T) \tag{3}$$

其中，T表示token化后的文本输入，h代表融合Transformer输出的最终多模态上下文表示。通过跨模态的注意力机制，视觉信息能够有效融入语言模型的特征表示中。这一方式已在多个领域取得了显著进展，如Flamingo[40]在大语言模型的冻结Transformer层之间插入额外的交叉注意力层，从而通过外部视觉线索增强语言特征，而CogVLM[70]在每个Transformer层中插入了一个视觉专家模块，以实现视觉和语言特征之间的双向交互和融合，然而，这种多模态特征投影器在RS-MLLM中的应用仍有待进一步探索和发展。

以上3种多模态特征投影器因自身结构特点，适用于不同的应用场景。基于投影的多模态投影器采用简单MLP，轻量化且参数量小，并且在训练过程中由于独立在多模态编码器与大语言模型之外，可以冻结多模态编码器与大语言模型进行训练，适合资源受限的场景；而基于查询的多模态特征投影器通过更深度的特征融合，能够提升模型的性能表现，并且由于同样独立于多模态编码器与大语言模型之外，同样可以冻结多模态编码器与大语言模型进行训练，减小训练成本，适用于对精度要求较高并具有训练资源充足的场景；而基于融合的多模态特征投影器通过更深度特征融合以期获得最佳表现，但由于改变了大语言模型本身的结构，需要重新对大语言模型进行训练，适用于对精度要求极高且能够承受更高训练成本的场景。

3.3 预训练大语言模型

预训练大语言模型普遍采用Transformer[25]架构，具备强大的上下文理解与文本生成能力。其核心设计包括多头自注意力机制、多层解码器堆叠、残差连接、归一化(如 LayerNorm 或 RMSNorm)以及位置编码(如绝对或旋转位置编码)，使模型能够高效捕捉长距离依赖与语义关系。在训练过程中，通常采用自回归语言建模作为训练目标，通过在大规模语料(如网络文本、百科、对话数据等)上进行自监督学习，使模型逐步掌握语言结构与语义知识。训练阶段还结合大规模分布式计算框架，采

用数据并行、模型并行及混合精度训练等策略，以支持百亿级参数模型的高效优化。得益于上述结构与训练机制，预训练大语言模型成为RS-MLLM中的核心模块，能够在多模态特征投影器的支持下，实现跨模态信息的融合、理解与自然语言生成。考虑到训练成本和易获取性等，现有的RS-MLLM普遍采用开源的预训练大语言模型作为基础模块，主要包括LLaMA和Vicuna等。

LLaMA[22]是由Meta开发的大规模预训练语言模型，与GPT-2等Transformer模型类似，但进行了多项改进。包括：将激活函数从 ReLU 改为SwiGLU，使用旋转位置编码(RoPE)替代绝对位置编码，使用均方根归一化(RMSNorm)替代标准的层归一化(LayerNorm)。这些优化旨在在保持高效架构的前提下提升模型性能，使其更适用于各类自然语言处理任务。LLaMA通过轻量化的设计和高参数效率，特别适用于资源有限的环境，在文本生成、机器翻译、机器问答等任务中表现突出。此外，LLaMA的开源特性为研究人员和开发者提供了灵活的模型优化和微调空间。RS-CapRet[45], LHRS-Bot[11]和Popeye[47]等RS-MLLMs均以LLaMA作为基础模块。

Vicuna[38]是在 LLaMA 基础上针对对话场景优化得到的版本，通过引入高质量对话数据进行指令微调，显著提升了模型的上下文理解和对话生成能力。Vicuna专注于对话系统场景，能够更自然地处理多轮对话、回答复杂问题和生成流畅的交互式文本。相比通用语言模型，Vicuna在对话任务中的表现更加优越，同时继承了LLaMA的高效性和开放性。CDChat[49], RSGPT[10]和RS-LLaVA[64]等RS-MLLMs均使用Vicuna作为基础模型。

4 模型训练

在遥感多模态大语言模型的训练过程中，不同阶段对数据的需求呈现出显著差异。当前多模态大语言模型主流训练范式多采用“三阶段策略”，包括跨模态对齐(Alignment)[71]、指令微调(Instruction Tuning)[26]与人类反馈强化学习(Reinforcement Learning with Human Feedback, RLHF)[72]。与之相对应，所使用的数据集大致可分为3类：跨模态对齐数据集、指令微调数据集与人类反馈强化学习数据集。它们在结构、构建方式与语义覆盖范围上各具特点，共同决定了模型对遥感任务的理解深度与泛化能力。本节将对模型不同阶段训练方法与训练数据集进行介绍。

4.1 训练方法

(1) 跨模态对齐：冻结多模态编码器与预训练的大语言模型，并利用大量数据集对随机初始化的多模态特征投影器进行微调，从而在对齐多模态信息的同时减少需要训练的参数。而常见的对齐方式包括图文对比(ITC)、图文匹配(ITM)以及图像引导的文本生成(ITG)。当前遥感多模态大语言模型普遍采用基于 ITG 的对齐策略，即仅使用语言建模损失(CrossEntropy)对输出文本进行监督，从而实现视觉语义向语言空间的有效映射。该训练方式在结构上与后续的指令调优阶段高度一致，仅在优化目标与可训练模块上存在差异。

(2) 指令调优：在该阶段，通常冻结已训练的多模态特征投影器和多模态编码器，结合自行构建的指令调优数据集对预训练大语言模型进行微调，从而使得模型能够理解各种任务，并提升模型在相应任务上的性能。其中，为了减少所需训练的参数，该过程通常采用低秩适配(Low-Rank Adaptation，LoRA)[73]方法对大语言模型进行微调。典型的指令调优样本通常包括可选择的指令和输入输出对，指令是通过自然语言描绘的任务含义，例如“Describe the image in detail.”，输入则根据具体任务进行选择，例如在遥感图像描述任务中只需要输入图像，而在遥感图像问答任务中则需要输入图像文本对，输出则是遵循指令并根据输入情况得到的回答。跨模态指令样本可以表示为3元组形式，即$(\mathcal{I},\mathcal{M},\mathcal{R})$，其中$\mathcal{I},\mathcal{M},\mathcal{R}$分别代表指令、跨模态输入和真实响应。模型根据指令和跨模态输入预测答案：

$$\mathcal{A}=f\left(\mathcal{I},\mathcal{M};\theta\right) \tag{4}$$

其中，$\mathcal{A}$表示预测的答案，θ是模型的参数。训练目标通常是指用于训练大语言模型的原始自回归目标，基于该目标，模型被鼓励预测响应的下一个词。该目标可以表示为

$$\mathcal{L}\left(\theta\right)=-\sum_{i=1}^{N}\log p\left(\mathcal{R}_i\mid\mathcal{I},\mathcal{R}_{<i};\theta\right) \tag{5}$$

其中，N是真实响应的长度，$p\left(\mathcal{R}_i\mid\mathcal{I},\mathcal{R}_{<i};\theta\right)$表示在给定上下文条件下，第$i$个token$\mathcal{R}_i$出现的概率。

(3) 人类反馈强化学习：RLHF是一种将人类偏好引入模型优化过程的训练范式，旨在增强生成模型对复杂、多样人类指令的理解和响应能力。其核心流程包括3阶段：首先，通过人工标注或比较生成结果的优劣构建奖励模型(Reward Model, RM)；其次，冻结语言模型主体，利用强化学习算法(如

Proximal Policy Optimization, PPO)对模型行为进行优化，以最大化奖励模型输出的得分；最后，通过人类反馈引导模型生成更加符合用户意图的响应。在多模态大语言模型中，RLHF 常作为指令调优后的进一步精调手段，用以提升模型在开放式问答、图文理解等任务中的交互性与安全性。与前期对齐或监督微调阶段相比，RLHF 在训练目标上更关注用户偏好而非单一任务正确性，从而实现更自然、更有用的模型输出。

本文针对不同模型的训练阶段也进行了系统梳理，大多数RS-MLLMs多采用指令调优和多模态对齐两个阶段进行训练，而人类反馈强化学习在该领域的应用尚处于探索阶段。由于具体实现方式存在差异，不同模型在训练策略上也有所不同。例如，ChangeChat[48], GeoChat[42], TEOChat[50]和SkySenseGPT[44]是直接加载相关研究[71]中已训练好的多模态特征投影器，因此上述模型只需进行指令调优。而EarthMarker[61], LHRS-Bot[11]等模型在跨模态对齐以及指令调优阶段之间，额外引入了针对大语言模型的直接微调阶段，从而提升了模型在特定任务中的适应能力与性能。

4.2 训练数据集

(1) 跨模态对齐数据集：在跨模态对齐阶段，数据集主要承担基础模态对齐与语义映射的功能。该阶段的核心目标是通过大规模遥感图文对样本，学习视觉特征到语言表示空间之间的映射关系，从而实现不同模态间的有效对齐。该类数据通常采用结构化的键值对格式，如：

{"image":⋯,"caption":⋯}或{"image":⋯,"question":⋯,"answer":⋯}

在大规模数据集中，这些格式常配合WebDataset或Parquet格式使用，以提升数据加载效率和并行处理能力[12]。

随着样本量的提升，模型在图文对齐和语义泛化方面的能力表现出明显增长。例如，数十万规模的数据集适用于基本语义转换与模态映射任务，而百万级以上的数据集(如EarthGPT[12], RSVP-3M[61])则能有效支撑更复杂的多样化任务(如语义分割到问答，检测到区域描述的转换)，并能在跨区域、跨时间规模上提升泛化能力。同时，大规模数据还能有效地缓解长尾类别问题，使模型对边缘语义具备更高识别鲁棒性[11,12,61]。

跨模态对齐数据通常依托多个异构遥感数据源，采用模板化改写、弱标签翻译、外部知识添加(如OpenStreetMap, Segment Anything Model)和大语言模型(如GPT-3/4)生成等策略，将原始结构化或半结构化遥感数据转化为符合自然语言表达方式的图文对样本。例如，LHRS-Align[11]借助OSM(OpenStreetMap)元数据自动生成图像描述，提升语义一致性与对齐效率；RSVP-3M[61]结合SAM区域提示与GPT模型生成区域级文本，增强模型对局部语义的表达能力；EarthGPT[12]则统一将遥感分类、检测与定位任务转化为单轮问答格式，实现多任务对齐结构的标准化。

(2) 指令调优数据集：指令微调阶段则更加关注模型基于自然语言指令完成图像描述、目标定位、变化检测、问答推理等复杂任务。该阶段数据集包含多种结构，如Alpha结构：

{"input":{"image":⋯},"instruction":"⋯","output":"⋯"}

或类LLaVA对话式结构[26,31,44]：

{"image":⋯,"conversations":[{"from":"human","value":"⋯"},{"from":"assistant","value""⋯"}]}

同时，为提升空间推理能力，指令或输入中常带有(边界框)、(区域遮罩)、(地理坐标)等结构化token，以明确视觉关注区域[46,47]。

现有的RS-MLLM通常采用“三步式策略”构建指令微调数据：第1步，从现有遥感任务数据中提取关键标注信息(如分类标签、目标框、变化区域等)，通过模板重写将其转化为自然语言形式的输入输出对；第2步，借助GPT类语言模型生成高质量的指令响应内容，模拟多轮对话或复杂推理过程；第3步，引入人工审核与语义过滤机制，以提升样本的准确性、逻辑性和语义多样性[31,61]。

在这一流程下，不同模型在指令样本构建策略上展现出多样化进化路径。GeoChat[42]于2023年底提出，通过围绕目标检测框生成多轮问答，有效构建图像区域的语义表征。SkyEyeGPT[31]于2024年初发布，针对近百万条图文样本进行了人工筛查，确保数据质量，并将图像描述与问答任务统一封装为标准化指令格式，为多任务统一学习提供基础。LHRS-Instruct[11]于2024年2月提出，显式融合地理坐标点位与OSM元数据，构建具备地理空间推理能力的复杂问答指令，显著增强模型的空间理解能力。FIT-RS[44]于2024年6月推出，系统构建了百万级细粒度图文任务指令数据集，并在图像描述、问答和关系推理等任务中实现标准封装与任务统一。EarthMarker[61]于2024年7月提出，通过在输入中引入区域与地理点位提示，增强模型对空间显著性区域的感知能力。IFShip[46]则于2024年8月在构建过程中引入遥感领域专家知识，设计步骤式链式推

理任务，提升模型对专业概念的理解与多步逻辑建模能力。GeoGround[47]在2024年11月提出，结合目标检测框生成多轮区域问答，进一步拓展对局部区域语义细化理解的能力。

(3) 人类反馈强化学习数据集：人类反馈强化学习作为近年来大语言模型领域的重要训练范式，为多模态模型生成质量与响应可信度的提升提供了新思路。在遥感多模态大语言模型研究中，RLHF仍处于探索起步阶段，其核心思想是通过人类偏好引导模型优化输出效果。RLHF 训练通常构建包含多个候选响应的样本集合，并由人工标注或偏好模型判断响应质量，生成用于训练奖励模型的数据格式，如：

{"prompt":⋯, "response_A":⋯, "response_B":⋯, "preference":⋯}

受限于遥感任务的专业复杂性与高标注成本，目前尚缺乏公开的遥感领域 RLHF 数据集，但该方向在提升遥感模型的生成可靠性、专家适配性与应用价值方面具有重要发展潜力。

本文统计了现有RS-MLLMs的训练数据集，由于指令调优是目前RS-MLLM研究更为大家关注也是使用最为广泛的阶段，因此重点整理了RS-MLLM训练时使用的指令微调数据集，如表2所示。可以看出，现有的RS-MLLM研究更倾向于数据驱动，而非模型驱动，研究重点更多集中在指令调优数据集的调整与优化。

表 2 遥感多模态大语言模型训练时使用的数据集

Tab. 2 Datasets used for train in RS-MLLMs

RS-MLLM	指令调优	其他训练
AeroLite[57]	-	RSSCN7, DLRSD, iSAID, LoveDA, WHU, UCM-Captions, Sydney-Captions
Aquila[55]	FIT-RS	CapERA, UCM-Captions, Sydney-Captions, NWPU-Captions, RSICD, RSITMD, RSVQA-HR, RSVQA-LR, WHU_RS19
Aquila-plus[56]	Aquila-plus-100K	-
CDChat[49]	LEVIR-CD, SYSU-CD	LEVIR-CD, SYSU-CD
ChangeChat[48]	ChangeChat-87k	-
EagleVision[58]	EVAttrs-95K	EVAttrs-95K
EarthDial[59]	EarthDial-Instruct	EarthDial-Instruct
EarthGPT[12]	MMRS-1M	LAION-400M, COCO Caption
EarthGPT-X[60]	M-RSVP	-
EarthMarker[61]	RSVP-3M	COCO Caption, RSVP-3M, RefCOCO, RefCOCO+
GeoChat[42]	RS multimodal instruction following dataset	-
GeoGround[29]	refGeo	FAIR1M, DIOR, DOTA
GeoLLaVA-8K[62]	SuperRS-VQA, HighRS-VQA	-
IFShip[46]	TITANIC-FGS	-
LHRS-Bot[11]	LLaVA complex reasoning dataset, NWPU, RSITMD, LHRS-Instruct	LHRS-Align
LHRS-Bot-Nova[63]	multi-task instruction dataset	LHRS-Align-Recap, LHRS-Instruct, LHRS-Instruct-Plus, LRV-Instruct
Popeye[47]	MMShip	COCO Caption
RingMoGPT[30]	instruction-tuning dataset	image-text pre-training dataset
RS-CapRet[45]	-	RSCID, UCM-Captions, Sydney-Captions, NWPU-Captions
RSGPT[10]	-	RSICap
RS-LLaVA[64]	RS-Instructions	-
RSUniVLM[65]	RSUniVLM-Instruct-1.2M	RSUniVLM-Resampled
SkyEyeGPT[31]	SkyEye-968k	SkyEye-968k
SkySenseGPT[44]	FIT-RS, NWPU-Captions, UCM-Captions, RSITMD, EarthVQA, Floodnet-VQA, RSVQA-LR, DOTA, DIOR, FAIR1M	-
Spectral-LLaVA[66]	BigEarthNet-v2	fMoW, BigEarthNet-v1
TEOChat[50]	TEOChatlas	-
UniRS[13]	GeoChat-Instruct, LEVIR-CC, EAR	-
VHM[43]	VersaD-Instruct, VariousRS-Instruct, HnstD	VersaD

注：由于主要关注指令调优阶段的数据集，因此将其他训练阶段的数据集合并至第3列。斜体表示该数据集在论文中并未给出正式名称或缩写。

5　多模态遥感解译任务及基准评测

随着RS-MLLM的快速发展，遥感任务正从单一模态分析逐步拓展到多模态融合，涵盖遥感影像、遥感视频、文本等多类型数据的综合处理。目前，遥感领域的核心任务包括遥感图像描述生成、遥感视觉问答、遥感视觉定位、遥感场景分类、遥感目标检测、遥感变化检测与描述，以及遥感图像检索等。由第4节的分析可知，相关任务数据集的构建是当前RS-MLLM研究的关键，本文系统梳理了RS-MLLM涉及的各项任务及其对应的基准数据集(见表3)，并综合评测典型遥感多模态大语言模型在多模态遥感解译任务中的表现。详细论述见下文。

表 3　遥感多模态大语言模型的适用任务及其对应的基准数据集

Tab. 3　The applicable tasks of RS-MLLMs and their corresponding benchmark datasets

RS-MLLM	RSIC	RSVQA	RSVG	RSSC
AeroLite[57]	Sydney-Captions[74], UCM-Captions[74]	–	–	–
Aquila[55]	RSICD[75], Sydney-Captions[74], UCM-Captions[74], FIT-RS[44]	RSVQA-LR[76], RSVQA-HR[76], FIT-RS[44]	–	–
EarthDial[59]	NWPU-Captions[77], RSICD[75], RSITMD-Captions[78], Sydney-Captions[74], UCM-Captions[74]	RSVQA-LR[76], RSVQA-HR[76]	–	AID[79], UCMerced[80], WHU-RS19[81], BigEarthNet[82], xBD Set 1[83], fMoW[84]
EarthGPT[12]	NWPU-Captions[77]	CRSVQA[85], RSVQA-HR[76]	DIOR-RSVG[86]	NWPU-RESISC45[87], CLRS[88], NaSC-TG2[89]
GeoChat[42]	–	RSVQA-LR[76], RSVQA-HR[76]	GeoChat*[42]	AID[79], UCMerced[80]
GeoGround[29]	–	–	DIOR-RSVG[86], RSVG[90], GeoChat*[42], VRSBench*[91], AVVG[29]	–
LHRS-Bot[11]	–	RSVQA-LR[76], RSVQA-HR[76]	DIOR-RSVG[86], RSVG[90]	AID[79], WHU-RS19[81], NWPU-RESISC45[87], SIRI-WHU[92], EuroSAT[93], METER-ML[94], fMoW[84]
LHRS-Bot-Nova[63]	–	RSVQA-LR[76], RSVQA-HR[76]	DIOR-RSVG[86], RSVG[90]	AID[79], WHU-RS19[81], NWPU-RESISC45[87], SIRI-WHU[92], EuroSAT[93], METER-ML[94], fMoW[84]
RingMoGPT[30]	DOTA-Cap[30], DIOR-Cap[30], NWPU-Captions[77], RSICD[75], Sydney-Captions[74], UCM-Captions[74]	HRVQA[95]	–	AID[79], NWPU-RESISC45[87], UCMerced[80], WHU-RS19[81]
RS-CapRet[45]	NWPU-Captions[77], RSICD[75], Sydney-Captions[74], UCM-Captions[74]	–	–	–
RSGPT[10]	RSIEval[10], UCM-Captions[74], Sydney-Captions[74], RSICD[75]	RSIEval[10], RSVQA-LR[76], RSVQA-HR[76]	–	–
RS-LLaVA[64]	UCM-Captions[74], UAV[96]	RSVQA-LR[76], RSIVQA-DOTA[97]	–	–
RSUniVLM[65]	–	RSVQA-LR[76], RSVQA-HR[76]	DIOR-RSVG[86], VRSBench[91]	AID[79], WHU-RS19[81], NWPU-RESISC45[87], SIRI-WHU[92]
SkyEyeGPT[31]	UCM-Captions[74], CapERA[98]	RSVQA-LR[76], RSVQA-HR[76]	DIOR-RSVG[86], RSVG[90]	–
TEOChat[50]	–	RSVQA-LR[76], RSVQA-HR[76]	–	AID[79], UCMerced[80]
UniRS[13]	–	RSVQA-LR[76], RSVQA-HR[76], CRSVQA[85]	–	–
VHM[43]	–	RSVQA-LR[76], RSVQA-HR[76]	DIOR-RSVG[86]	AID[79], WHU-RS19[81], NWPU-RESISC45[87], SIRI-WHU[92], METER-ML[94]

注：*表示测试集已修改。

5.1 模型适用任务

目前，大多数RS-MLLM主要聚焦于遥感图像描述生成(RSIC)、遥感视觉问答(RSVQA)、遥感视觉定位(RSVG)以及遥感场景分类(RSSC)4项典型任务。为帮助厘清不同模型所适配的任务类型，表3汇总了现有模型在各任务上的适用性，并列出了其在原始文献中所采用的评估数据集。在RSIC任务中，已用于该任务的模型有AeroLite[57], Aquila[55], EarthDial[59], EarthGPT[12], EarthMarker[61], RingMoGPT[30], RS-CapRet[45], RSGPT[10], RS-LLaVA[64], SkyEyeGPT[31]和SkySenseGPT[44]，评估常基于RSICD[75], NWPU-Captions[77], Sydney-Captions[74]和UCM-Captions[74]等数据集进行。在RSVQA任务中，应用模型为Aquila[55], EarthDial[59], EarthGPT[12], GeoChat[42], LHRS-Bot系列[11,64]、RingMoGPT[30], RSGPT[10], RS-LLaVA[64], RSUni-VLM[65], SkyEyeGPT[31], SkySenseGPT[44], TEOChat[50], UniRS[13]和VHM[43]，常用数据集包括RSVQA-LR[76], RSVQA-HR[76]和CRSVQA[85]等。在RSVG任务中，涉及的模型有EarthGPT[12], GeoChat[42], GeoGround[29], LHRS-Bot系列[11,64]、RSUniVLM[65], SkyEyeGPT[31]和VHM[43]，主要采用DIOR-RSVG[86]和RSVG[90]等数据集进行测试。在RSSC任务方面，相关模型为EarthDial[59], EarthGPT[12], EarthMarker[61], GeoChat[42], LHRS-Bot系列[11,64]、RingMoGPT[30], RSUniVLM[65], SkySenseGPT[44], TEOChat[50]和VHM[43]，依托的数据集包括AID[79], UCMerced[80]和NWPU-RESISC45[87]等。此外，还有一些模型，如CDChat[49], ChangeChat[48], IFShip[46]和Popeye[47]，则主要针对变化检测、舰船识别等其他遥感任务进行设计与评估。各任务的具体内容与其典型数据集在下文中作详细介绍。

5.2 遥感图像描述任务

遥感图像描述任务是指模型通过分析遥感图像生成自然语言文本描述。RSIC任务常用的数据集包括RSICD[75]和NWPU-Captions[77]等，前者包含10,921张遥感图像，每张图像配有5条高质量文本描述，涵盖城市区域、自然景观等地理场景，是经典的RSIC任务数据集；后者则提供31,500张图像及157,500条文本描述，覆盖复杂场景和多样化语义，适合训练高性能模型。另外，Sydney-Captions[74]和UCM-Captions[74]等数据集有时也会作为辅助数据源，上述两个数据集分别提供了613张遥感影像(7个场景)和2,100张遥感图像(21个场景)，并附带丰富的文本描述。此外，RSGPT[10]则提出了一种新的基准数据集RSIEval，其包含100对图像描述对和936对视觉问答对，为评估RS-MLLM在RSIC任务和遥感视觉问答任务上的表现提供了基准。

表4展示了现有的RS-MLLM在不同的遥感图像描述数据集上的性能表现。可以看到，各模型在不同指标上的表现差异明显。例如，EarthDial[59]在NWPU-Captions数据集上取得了最高的METEOR分数(0.806)，表明其在生成语义一致、内容贴合的描述方面表现尤为出色，具备较强的图像理解与语义表达能力。然而，其在ROUGE-L指标上的表现相对较低，可能因为其采用了对话式语言建模架构，更强调语言表达的自然性和多样性，在句式结构和词序上不拘泥于参考描述。这种表达方式虽然增强了语义泛化能力，但在注重结构匹配的指标下得分相对受限。

相比之下，SkyEyeGPT[31]在多个数据集上BLEU系列得分均为最高或次高，在RSICD, UCM-Captions和Sydney-Captions 3个数据集上，其BLEU4分别达到了0.600, 0.784和0.774，展现出极强的语言表达一致性。该模型生成的描述在词汇选择与句法结构上更贴近人工标注，适用于需要标准化输出的场景。其优异表现得益于其所采用的 EVA-CLIP 多模态编码器，在捕捉遥感图像与语言描述之间细粒度语义对应关系方面具有卓越能力；同时，通过大规模指令微调进一步增强了其泛化能力，使模型在遥感图像描述任务中展现出稳健表现。

同样表现突出的还有AeroLite[57]模型，其在UCM-Captions和Sydney-Captions数据集上的BLEU-1, BLEU-4和ROUGE-L指标均达到了或接近最优水平，显示出极强的句法一致性和词汇精度，适合用于对语言形式严格要求的场景。其在句子结构和表达顺序方面表现稳健，是注重准确复述和结构匹配任务的优选模型。

此外，RingMoGPT[30]在遥感图像描述任务中展现出强劲的语义表达能力。该模型是在RingMo[99]和RingMoE[100]的基础上进一步发展的结果。RingMo提出了掩码图像建模(MIM)自监督学习范式，通过循环优化机制显著提升了从遥感图像中提取鲁棒的视觉语义特征的能力。RingMoE在RingMo的基础上引入了专家混合模型(MoE)技术，通过动态路由机制扩展了模型的容量，并提高了其灵活性和自适应性。与这两个模型侧重于视觉理解不同，RingMoGPT专注于将图像特征映射到文本空间，并通过LLM进行理解。可以看到RingMoGPT在遥感图像描述任务CIDEr指标上全面领先，在RSICD, UCM-

表 4 遥感多模态大语言模型在不同的遥感图像描述数据集上的性能表现

Tab. 4 Performance of RS-MLLMs on various remote sensing image captioning datasets

数据集	模型	BLEU-1 ↑	BLEU-2 ↑	BLEU-3 ↑	BLEU-4 ↑	METEOR ↑	ROUGE-L ↑	CIDEr ↑	SPICE ↑
NWPU-Captions [77]	EarthDial[59]	–	–	–	–	**0.806**	0.400	–	–
	EarthGPT[12]	**0.871**	**0.787**	0.716	0.655	0.445	**0.782**	1.926	**0.322**
	EarthMarker[61]	0.844	0.731	0.629	0.543	0.375	0.700	1.629	0.268
	RS-CapRet[45]	**0.871**	**0.787**	**0.717**	**0.656**	0.436	0.776	**1.929**	0.311
RSICD[75]	Aquila[55]	0.746	–	–	–	–	–	–	–
	EarthDial[59]	–	–	–	–	**0.562**	0.276	–	–
	RS-CapRet[45]	0.741	0.622	0.529	0.455	0.376	**0.649**	2.605	**0.484**
	RSGPT[10]	0.703	0.542	0.440	0.368	0.301	0.533	1.029	–
	SkyEyeGPT[31]	**0.867**	**0.767**	**0.673**	**0.600**	0.354	0.626	0.837	–
	RingMoGPT[30]	–	–	–	–	0.343	0.616	**2.758**	–
UCM-Captions [74]	AeroLite[57]	**0.934**	–	–	**0.796**	0.498	**0.880**	–	–
	Aquila[55]	0.883	–	–	–	–	–	–	–
	EarthDial[59]	–	–	–	–	**0.514**	0.342	–	–
	RS-CapRet[45]	0.843	0.779	0.722	0.670	0.472	0.817	3.548	**0.525**
	RSGPT[10]	0.861	0.791	0.723	0.657	0.422	0.783	3.332	–
	SkyEyeGPT[31]	0.907	**0.857**	**0.816**	0.784	0.462	0.795	2.368	–
	RS-LLaVA[64]	0.900	0.849	0.803	0.760	0.492	0.858	3.556	–
	RingMoGPT[30]	–	–	–	–	0.499	0.833	**3.593**	–
Sydney-Captions [74]	AeroLite[57]	**0.919**	–	–	0.759	0.475	**0.837**	–	–
	Aquila[55]	0.834	–	–	–	–	–	–	–
	EarthDial[59]	–	–	–	–	**0.573**	0.410	–	–
	RS-CapRet[45]	0.787	0.700	0.628	0.564	0.388	0.707	2.392	**0.434**
	RSGPT[10]	0.823	0.753	0.686	0.622	0.414	0.748	2.731	–
	SkyEyeGPT[31]	**0.919**	**0.856**	**0.809**	**0.774**	0.466	0.777	1.811	–
	RingMoGPT[30]	–	–	–	–	0.421	0.734	**2.888**	–
FIT-RS[44]	Aquila[55]	**0.351**	–	–	–	–	–	–	–
	GeoChat[42]	0.088	–	–	–	–	–	–	–
	SkySenseGPT[44]	0.273	–	–	–	–	–	–	–

注：所有结果均引自对应论文原文，加粗表示最佳结果。

Captions和Sydney-Captions 3个数据集上，RingMoGPT分别取得了2.758, 3.595和2.888的CIDEr分数，相较于第2名的2.605, 3.556和2.731分别高出0.153, 0.039和0.157，说明其生成的文本在与参考答案的语义接近度和多样性方面具备明显优势。虽然该模型在BLEU等精度导向指标上相对较弱，但CIDEr的强势表现反映出它在生成丰富、多样且具有实际参考价值的语言描述方面具有显著潜力，适合于强调语义表现力和人类可读性的遥感文本生成任务。

值得注意的是，由于RSICD数据集覆盖复杂场景和多样化语义的特点，模型在RSICD上的性能表现普遍不如在其他数据集上的效果。例如，在RSICD数据集上，SkyEyeGPT[31]的BLEU-4得分为0.600，低于UCM-Captions和Sydney-Captions，RS-CapRet[45]也表现得较为平平，得分为0.455。这可能是因为RSICD数据集具有更多复杂和多样化的场景，导致模型在细粒度语义映射方面面临更大的挑战，未来在该数据集上的性能仍有较大的提升空间。

5.3 遥感视觉问答任务

遥感视觉问答任务是指模型结合遥感图像与自然语言问题生成精准答案。其相关的数据集包括RSVQA-LR[76]和RSVQA-HR[76]。RSVQA-LR以低分辨率遥感图像为主，主要用于评估模型在资源受限或传感器能力有限条件下对图像语义和空间特征

的理解能力；RSVQA-HR则包含高分辨率遥感图像，聚焦于细粒度的特征分析和语义推理，适用于精细场景理解任务，如城市规划和灾害监测。HR-VQA[95]同样基于高分辨率遥感图像，涵盖更丰富的问题类型，如对象识别、数量计数、空间关系分析以及更深层次的语义理解，强调模型在复杂视觉推理任务中的表现。此外，CRSVQA[85]数据集引入了跨区域和跨语境的遥感图像问答设定，致力于提升模型在地理分布广泛、场景差异明显的遥感环境中的泛化能力。该数据集的问题设计更加多样，结合了图像变化检测、区域对比、复杂空间关系等任务，有助于推动遥感VQA模型朝着更强的迁移性和实用性方向发展。

表5展示了现有的RS-MLLM在不同的遥感视觉问答数据集上的性能表现。在RSVQA-LR数据集中，多数模型在不同评估指标(Presence、Compare、Rural/Urban)上的表现相对接近，但在单个维度上仍有明显优势模型。例如，SkySenseGPT在Presence指标中取得95.00%的最高准确率，UniRS在Compare指标上达到93.23%，RS-LLaVA在Rural/Urban分类任务中表现最佳(95.00%)。这表明在低分辨率图像条件下，模型通过大量数据集的训练，已具备较强的语言理解与基础图像判断能力，但不同模型在特定类型问题上仍存在结构性优势。从平均准确率来看，UniRS(92.63%), RSGPT(92.29%)与SkySenseGPT(92.69%)在多个指标上保持高水准，体现出良好的综合适应能力。

在RSVQA-HR数据集中，图像分辨率的提高进一步拉开了模型间在图像细节识别和复杂语义推理能力方面的差距。Aquila[55]凭借其SFI模块的精确特征对齐技术，在Presence指标中以92.64%的准确率领先，显现了强大的结构感知能力；LHRS-Bot[11]在Compare指标中达到92.53%，并在平均准确率上达到92.55%，表现出稳定的整体性能；LHRS-Bot-Nova[63]的平均成绩为92.06%，在高分辨率条件下也维持了较高准确率。LHRS-Bot系列[11,64]在多个维度上保持了高性能，尤其在高分辨率图像条件下展现了较强的优势，这得益于其多任务指令调优机制。由于采用了不同的多模态编码器(CLIP-ViT和SigLIP)，这两个模型在不同数据集和评估指标上展现了各自的结构性优势。

总体来看，Aquila[55], LHRS-Bot系列[11,64]和RSGPT[10]在不同分辨率条件下均表现出色，证明了这些模型在跨尺度图像处理和多类型视觉问答任务中的强大泛化能力与鲁棒性。

5.4 遥感视觉定位任务

遥感视觉定位任务是指模型根据自然语言描述，定位遥感图像中的具体目标区域(如地标或建筑物)。RSVG[90]和DIOR-RSVG[86]数据集是用于此类任务

表 5 遥感多模态大语言模型在不同的遥感视觉问答数据集上的性能表现

Tab. 5 Performance of RS-MLLMs on various remote sensing visual question answering datasets

模型	数据集								
	RSVQA-LR[76]				RSVQA-HR[76]			CRSVQA[85]	FIT-RS[44]
	Presence(%)	Compare(%)	Rural/Urban(%)	Avg(%)	Presence(%)	Compare(%)	Avg(%)	Avg(%)	Avg(%)
Aquila[55]	92.72	–	–	–	**92.64**	–	–	–	**83.87**
EarthDial[59]	92.58	92.75	94.00	**92.70**	58.89	83.11	72.45	–	–
EarthGPT[12]	–	–	–	–	62.77	79.53	72.06	82.00	–
GeoChat[42]	91.09	90.33	94.00	90.70	59.02	83.16	–	–	53.47
LHRS-Bot[11]	89.07	88.51	90.00	89.19	92.57	**92.53**	**92.55**	–	–
LHRS-Bot-Nova[63]	89.00	90.71	89.11	89.61	91.68	92.44	92.06	–	–
RSGPT[10]	91.17	91.70	94.00	92.29	90.92	90.02	90.47	–	–
RS-LLaVA[64]	92.27	91.37	**95.00**	88.10	–	–	–	–	–
RSUniVLM[65]	92.00	91.51	92.65	92.05	90.81	90.88	90.85	–	–
SkyEyeGPT[31]	88.93	88.63	75.00	84.19	80.00	80.13	82.56	–	–
SkySenseGPT[44]	**95.00**	91.07	92.00	92.69	69.14	84.14	76.64	–	79.76
TEOChat[50]	91.70	92.70	94.00	–	67.50	81.10	–	–	–
UniRS[13]	91.81	**93.23**	93.00	92.63	59.29	84.05	73.15	**86.67**	–
VHM[43]	91.17	89.89	88.00	89.33	64.00	83.50	73.75	–	

注：所有结果均引自对应论文原文，加粗表示最佳结果。

的基准数据集，RSVG数据集专注于遥感图像的视觉定位任务，旨在利用自然语言表达(例如“在图像中找到红色屋顶的建筑”)定位特定对象，覆盖城市、农村、森林、海岸等常用遥感场景。DIOR-RSVG则是基于DIOR数据集扩展而成的，包含图像、语言表达和目标边框3元组，涵盖更广泛的对象类别，包括车辆、建筑、自然景观等。

表6展示了现有的RS-MLLM在不同的遥感视觉定位数据集上的性能表现。从结果来看，LHRS-Bot-Nova[63]在DIOR-RSVG和RSVG两个数据集上均取得了最高的定位准确率(Pr@0.5分别为92.87%和81.85%)，相较第2名的88.59%以及73.45%提升了4.28和8.40个百分点，这得益于其采用的SigLIP编码器和多任务指令调优机制，有效增强了在复杂遥感场景中对语义目标进行精确定位的能力。而其余模型在两个数据集上均存在明显性能差距，在细粒度语义建模和高精度定位方面仍有提升空间。

同时可以注意到，由于RSVG数据集具有高图像分辨率、大幅宽的特点，其视觉定位任务的难度相较于DIOR-RSVG数据集更高。因此，模型在RSVG数据集上的表现普遍不如在DIOR-RSVG数据集上的效果。未来，模型需要优化大幅宽场景下视觉定位的精度，从而在RSVG数据集上取得更好的表现。

5.5 遥感场景分类任务

遥感场景分类任务是指模型通过遥感影像识别对应的地理场景类别，例如城市、森林或农业区等。RSSC任务常用的数据集包括NWPU-RESISC45[87]和AID[79]，前者包含31500张图像，覆盖45种场景类别，是场景分类的重要基准；后者包含10000张图像，涵盖30种场景类别，常用于高分辨率场景分类。此外，WHU-RS19[81]和UCMerced[80]数据集也被广泛应用，WHU-RS19包含19个场景类别，总计1005张图像，涵盖了城市、农业、森林等多种地理环境，适用于多尺度和多分辨率的场景分类研究；UCMerced数据集则包含21个类别，共有2100张高分辨率图像，主要用于评估高分辨率遥感图像分类算法的性能。

表 6　遥感多模态大语言模型在不同的遥感视觉定位数据集上的性能表现

Tab. 6　Performance of RS-MLLMs on various remote sensing visual grounding datasets

模型	DIOR-RSVG[86]	RSVG[90]
	Pr@0.5 (%)	Pr@0.5 (%)
EarthGPT[12]	76.65	–
GeoGround[29]	77.73	26.65
LHRS-Bot[11]	88.10	73.45
LHRS-Bot-Nova[63]	**92.87**	**81.85**
RSUniVLM[65]	72.47	–
SkyEyeGPT[31]	88.59	70.50
VHM[43]	56.17	–

注：所有结果均引自对应论文原文，加粗表示最佳结果。

表7展示了现有的RS-MLLM在不同遥感场景分类数据集上的性能表现。遥感场景分类数据集聚焦于地物分布、纹理模式和空间结构等宏观特征的识别，与传统目标检测任务侧重于局部实体定位的建模方式存在显著差异，因此对模型的视觉语义建模能力和跨模态融合能力提出更高要求。不同数据集的特性对模型能力构成了差异化要求。AID与NWPU-RESISC45类别数量多、来源广泛，强调多类区分能力；EuroSAT[93]与fMoW[84]引入多光谱和时序信息，对特征融合与跨尺度建模提出挑战；而SIRI-WHU[92], WHU-RS19[81]和NaSC-TG2[89]样本量相对较小，类别间差异较弱，更适合评估模型在小样本和语义重叠条件下的表现稳健性。

从已有测试结果来看，RingMoGPT[30]在AID (97.94%), NWPU-RESISC45 (96.47%)和WHU-RS19(97.71%)等数据集上表现突出，展现出良好的泛化能力和多场景适应性；LHRS-Bot-Nova[63]在SIRI-WHU (74.75%)和EuroSAT (63.54%)中取得优异成绩，说明其在面向细粒度本地化分类任务时具备一定的优势；EarthMaker[61]在UCMerced (86.52%)上也表现良好。整体来看，这些模型所采用的多模态编码器(EVA-CLIP, SigLIP和DINOv2)展现出结构性的优势，使得不同模型能够在面对不同分布特征的数据集时展现出各自的优势。而Earth-Dial[59]在UCMerced (92.42%)和fMoW (70.03%)上表现优秀，尤其在fMoW这类场景多变、语义模糊的数据集中仍保持较高准确率，体现出良好的鲁棒性和适应能力。这主要得益于其统一的多模态架构设计，结合了自适应高分辨率处理和多源数据融合，使模型在感知细节和建模复杂语义方面具备更强能力。

尽管部分模型已在大规模、多类别数据集上取得良好表现，但在小样本、语义重叠或跨尺度条件下仍表现出一定性能波动。例如，在SIRI-WHU和EuroSAT等数据集中，多数模型准确率显著低于其在AID或NWPU-RESISC45等数据集上的表现，反映出遥感场景分类任务仍具挑战性。未来工作可围绕更具结构适应性和尺度感知能力的视觉编码模

块设计，以及更具代表性、异质性和多层次语义覆盖的数据集构建方向，进一步提升模型的泛化性与鲁棒性。

5.6 其他任务与数据集

除了前述任务，遥感领域的其他任务也在蓬勃发展，例如变化描述[48]、目标检测[30]、图像检索[45]、定位描述[30]、船只检测[47]、时序场景分类[50]等，这些任务进一步拓展了RS-MLLM的应用范围。例如，遥感变化描述任务通过比较不同时间点的遥感图像，检测地物的变化并生成自然语言描述。该任务常用的数据集包括LEVIR-CD[101], SYSU-CD[102]和LEVIR-CC[103]。LEVIR-CD数据集包含637对遥感图像，标注建筑物变化区域；SYSU-CD提供20000对时间序列图像，涵盖多种地理区域；LEVIR-CC是LEVIR-CD的扩展，增加了自然语言描述。遥感目标检测任务核心则在于从遥感图像中识别并定位特定目标，如建筑物、飞机和道路等。该任务常用的数据集包括DOTA[104]和DIOR[105]等。DOTA数据集提供了高分辨率遥感图像及18个类别中1793658个目标的精确标注，是目标检测研究的核心资源；DIOR数据集则包含了20个类别，共计192472个遥感目标，适用于复杂场景下的目标检测任务。而遥感图像检索任务致力于实现基于自然语言文本检索遥感图像，或反向检索语义相符的文本描述。现有研究在遥感图像检索任务基准测试阶段通常使用RSIC任务中的相关数据集，例如，RS-CapRet[45]便在该任务中采用了RSICD[75]与UCM-Captions[74]作为其基准数据集。尽管这些任务对 RS-MLLM的发展具有重要意义，但现有模型较少在这些任务上实现相关功能，或开展系统性评测，相关能力仍有待进一步探索和验证。

近年来，RS-MLLMs在SAR影像处理任务中展现出逐步扩展的任务覆盖与语义理解能力。以EarthGPT[12]为例，其在SAR影像中已支持场景分类、目标检测、图像描述、视觉问答以及视觉定位等基础任务。而EarthDial[59]则在SAR任务上拓展了更广泛的任务类型与更高的细粒度水平，支持目标识别、语义分割、变化检测、图文问答与地理定位等多样任务，显著增强了模型对SAR数据的多任务适应能力和细粒度目标理解能力。

为支撑上述多样化的SAR任务，研究者围绕不同模型设计了大规模、高多样性的SAR影像数据集。EarthGPT[12]提出的MMRS 1M指令微调数据集整合了34个公开遥感数据集的光学、SAR与红外图像，形成超过100万对图文样本，涵盖城市、农田、水体、车辆、船只、飞机等多种目标，标注类型包括多标签分类、检测框(支持水平框与旋转框)以及图文配对等，充分丰富了模型的多模态训练语料。EarthDial[59]构建的EarthDial-Instruct数据集则在任务类型与数据规模上进一步扩展，包含超1111万组图文配对样本，重点引入Sentinel 1, xView3等主流SAR数据源，涵盖油轮、集装箱卡车、港口设施、建筑、道路等目标对象，标注形式覆盖多标签分类、带属性说明的检测框及自然语言问答对，显著提升了数据的语义细粒度与任务适配性。这些高

表 7 遥感多模态大语言模型在不同的遥感场景分类数据集上的性能表现

Tab. 7 Performance of RS-MLLMs on various remote sensing scene classification datasets

模型	UCMerced[80] (%)	AID[79] (%)	NWPU-RESISC45[87] (%)	CLRS[88] (%)	NaSC-TG2[89] (%)	WHU-RS19[81] (%)	SIRI-WHU[92] (%)	EuroSAT[93] (%)	METER-ML[94] (%)	fMoW[84] (%)
EarthDial[59]	**92.42**	88.76	–	–	–	96.21	–	–	–	**70.03**
EarthGPT[12]	–	–	93.84	**77.37**	**74.72**	–	–	–	–	–
EarthMaker[61]	86.52	77.97	–	–	–	–	–	–	–	–
GeoChat[42]	84.43	72.03	–	–	–	–	–	–	–	–
LHRS-Bot[11]	–	91.26	83.94	–	–	93.17	62.66	51.40	69.81	56.56
LHRS-Bot-Nova[63]	–	88.32	86.80	–	–	95.63	**74.75**	**63.54**	70.05	57.11
RingMoGPT[30]	86.48	**97.94**	**96.47**	–	–	**97.71**	–	–	–	–
RSUniVLM[65]	–	81.18	86.86	–	–	84.91	68.13	–	–	–
SkyEyeGPT[31]	60.95	26.30	–	–	–	–	–	–	–	–
TEOChat[50]	86.30	80.90	–	–	–	–	–	–	–	–
VHM[43]	–	91.70	94.54	–	–	95.80	70.88	–	**72.74**	–

注：所有结果均引自对应论文原文，加粗表示最佳结果。

质量、多样化的数据集为RS-MLLM在SAR场景下的多任务学习与跨模态遥感语义建模提供了坚实的数据支撑。

尽管如此，SAR图像在数据集构建中仍面临显著挑战。SAR采用主动微波成像，尽管具有全天时、全天候优势，但其影像呈现的是电磁散射特性，常伴随侧视畸变、散斑噪声与阴影效应，缺乏颜色与纹理等直观结构。这些物理机制导致非专业人员难以进行有效判读，显著增加了高质量数据标注的难度和成本。同时，由于SAR图像的空间分辨率通常低于光学图像，在小目标检测与细粒度语义理解中面临感知瓶颈。这些问题对RS-MLLM的精度与鲁棒性构成挑战，尤其在任务迁移与统一建模时表现突出。

为支持多样化任务的研究，近年来也涌现出一批多模态数据集，这些数据集主要是针对特定任务或用于评估现有模型的特定能力。例如，Hog[106]为遥感图像检索任务提供了新的数据支持；ChatEarthNet[107]结合了Sentinel-2卫星图像与ESA WorldCover的语义信息，具备全球覆盖范围的影像和对应的丰富文本描述；DDFAV[108]数据集则囊括了城市和农村地区的多种场景以及卫星和无人机的多样视角，可支持遥感图像描述、遥感视觉问答和复杂推理等任务。而部分数据集的出现填补了模型特定能力测试的空白，例如COREval[109]可用于评估遥感多模态大语言模型的感知与推理能力，特别适用于高层次推理任务，如遥感图像的视觉问答与推理；GeoMath[110]则聚焦于遥感图像的数学推理任务，旨在测试模型在遥感图像中的推理能力，推动了遥感领域数学推理的研究。

为了推动RS-MLLM的发展，数据集的构建是一个关键因素。然而，数据集的获取通常面临诸多挑战，尤其是在大规模标注和多模态数据生成方面。为了解决这些问题，RSTeller[111]开发了一个完善的数据集生成流程，包括原始数据获取、基于LLM的图像描述生成以及最终的数据集编译。该流程最终产生了超过100万张遥感图像，每张图像均配有多个描述性文本，此种数据集构建方式有效减少了人工标注的成本，也为大规模遥感图像的语义研究提供了新的思路。

6　拓展应用

RS-MLLM已在多项遥感任务中展现出显著的有效性与广泛的应用潜力，因此，部分研究者尝试将其能力进一步扩展，引入智能体范式构建遥感智能体系统，即在大语言模型的输出端集成外部工具，通过代理机制实现复杂遥感多模态任务的自动化处理与能力拓展。

TreeGPT[112]被认为是首个遥感智能体，其通过与外部工具进行交互，大幅提升了在多种遥感任务调度中的适用性。随后，Remote Sensing ChatGPT[113], Change-Agent[114]和RS-Agent[115]也相继出现，并在遥感场景下的特定任务中展现出了强大的执行能力。这些模型的发展时间线见图2。本文以RS-Agent 为代表，介绍其智能体系统架构、工具集成方式及当前所支持的遥感任务类型。

RS-Agent整体架构如图3所示，整个系统围绕大语言模型驱动的中央控制器展开，模拟“记忆—推理—反思”认知过程，实现对遥感任务的智能化处理。系统主要包括中央控制器(Central Controller)、解决方案空间(Solution Space)、知识空间(Knowledge Space)与工具空间(Tool Space)4个核心部分。中央控制器负责解析用户输入、规划任务流程并协调工具调用；解决方案空间为不同遥感任务提供结构化的处理方案指导；知识空间集成遥感领域的专业知识，为复杂推理与决策提供支持；工具空间集成多种遥感图像处理工具，支撑各类遥感任务的具体执行。RS-Agent通过中央控制器协调各模块协同工作，结合任务感知检索(TAR)和DualRAG机制，实现任务识别、方案检索、知识补充与工具调用的动态闭环，从而完成复杂遥感任务的端到端自动化处理。

工具集成能力在RS-Agent中发挥着关键作用，直接决定了系统应对复杂遥感任务的灵活性与适应性。面对遥感任务中多样化的场景类型与处理需求，单一模型往往难以胜任全部任务，需依赖多种算法工具的协同配合。为此，RS-Agent设计了灵活可扩展的工具空间，集成了包括目标检测、图像去噪、超分辨率、场景分类、图像描述等多种遥感图像处理工具。各工具通过FastAPI或函数接口方式与中央控制器集成，支持任务执行过程中的动态调度与组合调用。中央控制器依据任务规划结果自动选择最优工具序列，保障了任务处理的高效性与准确性。借助模块化工具架构，RS-Agent能够持续扩展工具库，适配新兴遥感任务需求，并可方便地集成到现有遥感数据处理平台中。

在多个公开遥感数据集的测试中，RS-Agent在场景分类、目标检测、目标计数与视觉问答任务上均取得优异成绩。其多轮任务规划与反思机制有效避免了多步骤任务中的累积误差，显著提升了长链条任务执行的稳定性与准确率。尤其在复杂任务如多类别目标检测与跨模态视觉问答中，RS-Agent

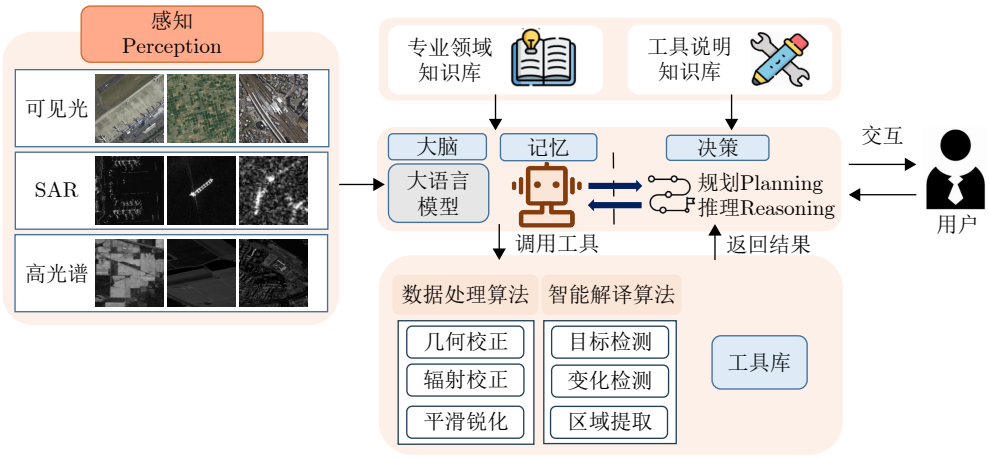

图 3 RS-Agent架构

Fig. 3 RS-Agent architecture

展现出较强的任务理解能力和动态调度能力，极大减少了人工干预和参数调整需求。

针对非光学遥感中的SAR图像处理，RS-Agent已经支持SAR目标检测与SAR飞机分类任务，表现出极强的适应性与扩展潜力。通过对SAR成像特性的理解与知识图谱支持，系统能够有效处理SAR图像中的散斑噪声、高对比边缘与结构化纹理等复杂特征，在目标区分与细分类任务中达到高精度。借助大语言模型的知识融合与推理能力，RS-Agent有望在后续拓展如SAR变化检测、海洋目标监测、干涉SAR(InSAR)分析等任务中，成为新一代非光学遥感智能分析的重要技术支撑。

尽管遥感智能体在多个任务中展现出良好性能，并具备跨模态调度能力，但其在复杂遥感环境中的实际部署仍面临诸多挑战与技术瓶颈。遥感数据本身具备强异构性，涵盖不同传感器类型(如光学、SAR、多光谱、热红外等)，如何实现对多模态信息的统一理解与动态融合仍是关键难点，例如，在SAR场景下，当前遥感智能体支持的任务类型仍较有限。未来，遥感智能体有望扩展所用工具，从而进一步支持SAR、多光谱等多模态任务，增强其在复杂遥感应用中的实用性与适应性。

同时，为推动遥感智能体的广泛应用和系统化评估，GeoLLM-Engine[116]提供了一个多任务平台，支持地理空间智能体执行复杂的地球观测任务。通过集成地理空间工具、动态地图和多模态知识库，该平台旨在提升智能体在遥感任务中的协作能力和推理水平；GeoLLM-QA[117] 则作为一套专为遥感与地理空间智能体设计的评估基准体系被提出，旨在为智能体性能的系统化量化评估提供标准依据。

7 总结与展望

近年来，遥感领域的多模态大语言模型取得了迅速发展，各类针对不同任务的模型被相继提出。本文聚焦遥感多模态大语言模型的发展，通过对其模型架构、训练方法、数据集、适用任务及其测评基准的分析，总结了遥感多模态大语言模型的研究进展和成果。

目前遥感多模态大语言模型的优化主要依赖于不同指令调优数据集的构建和训练，从而在特定任务上实现性能的显著提升，尽管已经取得了部分进展，但仍存在以下提升空间。

(1) 多源数据融合：目前遥感多模态大语言模型主要依赖光学图像与文本输入，难以处理SAR、红外、多光谱等遥感数据的异构性。针对这一问题，应引入成像机理与物理先验知识，研究SAR、红外、多光谱等多源遥感数据的融合与统一特征表示对齐方法，采集大规模多源数据集，实现多源异构信息的统一建模与语义对齐，从而提升模型对复杂遥感场景的综合感知能力。

(2) 多模态输出扩展：在灾害响应、军事侦察等遥感任务中，模型往往不仅需要生成文本描述，还需输出变化掩码、地物分类图、语音播报等多样化结果。为满足这些需求，RS-MLLM亟需拓展输出形式，深入研究多任务联合优化策略与多头解码机制，以提升其在实际场景中的应用能力和任务适应性。

(3) 模型可信性提升：随着RS-MLLM逐步应用于实际场景，其可信性问题日益受到关注。然而，在现有的 RS-MLLM 研究中，很少对模型的可信性进行专门的训练与测试。针对这一问题，需深入研究可信性训练方法，并设计科学的评估机制来测试模型的可信性。

(4) 轻量化模型研究：随着实际应用的推进，RS-MLLM有望部署在卫星上执行即时智能解译任

务。然而，由于星载平台计算资源通常受限，现有大部分模型难以直接部署。针对这一问题，未来应重点开展轻量化模型的设计与优化研究，以实现RS-MLLM在星上等资源受限环境中的高效运行与可靠应用。

利益冲突 所有作者均声明不存在利益冲突

Conflict of Interests The authors declare that there is no conflict of interests

参 考 文 献

[1] 王桥, 刘思含. 国家环境遥感监测体系研究与实现[J]. 遥感学报, 2016, 20(5): 1161–1169. doi: 10.11834/jrs.20166201.
WANG Qiao and LIU Sihan. Research and implementation of national environmental remote sensing monitoring system[J]. *Journal of Remote Sensing*, 2016, 20(5): 1161–1169. doi: 10.11834/jrs.20166201.

[2] 安立强, 张景发, MONTEIRO R, 等. 地震灾害损失评估与遥感技术现状和展望[J]. 遥感学报, 2024, 28(4): 860–884. doi: 10.11834/jrs.20232093.
AN Liqiang, ZHANG Jingfa, MONTEIRO R, *et al*. A review and prospective research of earthquake damage assessment and remote sensing[J]. *National Remote Sensing Bulletin*, 2024, 28(4): 860–884. doi: 10.11834/jrs.20232093.

[3] 张王菲, 陈尔学, 李增元, 等. 雷达遥感农业应用综述[J]. 雷达学报, 2020, 9(3): 444–461. doi: 10.12000/JR20051.
ZHANG Wangfei, CHEN Erxue, LI Zengyuan, *et al*. Review of applications of radar remote sensing in agriculture[J]. *Journal of Radars*, 2020, 9(3): 444–461. doi: 10.12000/JR20051.

[4] LI Yansheng, DANG Bo, ZHANG Yongjun, *et al*. Water body classification from high-resolution optical remote sensing imagery: Achievements and perspectives[J]. *ISPRS Journal of Photogrammetry and Remote Sensing*, 2022, 187: 306–327. doi: 10.1016/j.isprsjprs.2022.03.013.

[5] LI Yansheng, WEI Fanyi, ZHANG Yongjun, *et al*. HS^{2}P: Hierarchical spectral and structure-preserving fusion network for multimodal remote sensing image cloud and shadow removal[J]. *Information Fusion*, 2023, 94: 215–228. doi: 10.1016/j.inffus.2023.02.002.

[6] CHEN Yongqi, FENG Shou, ZHAO Chunhui, *et al*. High-resolution remote sensing image change detection based on Fourier feature interaction and multiscale perception[J]. *IEEE Transactions on Geoscience and Remote Sensing*, 2024, 62: 5539115. doi: 10.1109/TGRS.2024.3500073.

[7] 杨桄, 刘湘南. 遥感影像解译的研究现状和发展趋势[J]. 国土资源遥感, 2004(2): 7–10,15. doi: 10.3969/j.issn.1001-070X.2004.02.002.
YANG Guang and LIU Xiangnan. The present research condition and development trend of remotely sensed imagery interpretation[J]. *Remote Sensing for Land & Resources*, 2004(2): 7–10,15. doi: 10.3969/j.issn.1001-070X.2004.02.002.

[8] ZHAO W X, ZHOU Kun, LI Junyi, *et al*. A survey of large language models[J]. arXiv preprint arXiv: 2303.18223, 2023.

[9] YIN Shukang, FU Chaoyou, ZHAO Sirui, *et al*. A survey on multimodal large language models[J]. *National Science Review*, 2024, 11(12): nwae403. doi: 10.1093/nsr/nwae403.

[10] HU Yuan, YUAN Jianlong, WEN Congcong, *et al*. RSGPT: A remote sensing vision language model and benchmark[J]. *ISPRS Journal of Photogrammetry and Remote Sensing*, 2025, 224: 272–286. doi: 10.1016/j.isprsjprs.2025.03.028.

[11] MUHTAR D, LI Zhenshi, GU Feng, *et al*. LHRS-Bot: Empowering remote sensing with VGI-enhanced large multimodal language model[C]. The 18th European Conference on Computer Vision, Milan, Italy, 2024: 440–457. doi: 10.1007/978-3-031-72904-1_26.

[12] ZHANG Wei, CAI Miaoxin, ZHANG Tong, *et al*. EarthGPT: A universal multimodal large language model for multisensor image comprehension in remote sensing domain[J]. *IEEE Transactions on Geoscience and Remote Sensing*, 2024, 62: 5917820. doi: 10.1109/TGRS.2024.3409624.

[13] LI Yujie, XU Wenjia, LI Guangzuo, *et al*. UniRS: Unifying multi-temporal remote sensing tasks through vision language models[J]. arXiv preprint arXiv: 2412.20742, 2024.

[14] VOUTILAINEN A. A syntax-based part-of-speech analyser[C]. The 7th Conference of the European Chapter of the Association for Computational Linguistics, Dublin, Ireland, 1995.

[15] BRILL E and RESNIK P. A rule-based approach to prepositional phrase attachment disambiguation[C]. The 15th International Conference on Computational Linguistics, Kyoto, Japan, 1994.

[16] HINTON G E, OSINDERO S, and TEH Y W. A fast learning algorithm for deep belief nets[J]. *Neural Computation*, 2006, 18(7): 1527–1554. doi: 10.1162/neco.2006.18.7.1527.

[17] DEVLIN J, CHANG M W, LEE K, *et al*. BERT: Pre-training of deep bidirectional transformers for language understanding[C]. 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), Minneapolis, Minnesota, 2019: 4171–4186. doi: 10.18653/v1/N19-1423.

[18] RADFORD A, NARASIMHAN K, SALIMANS T, *et al*. Improving language understanding by generative pre-training[J]. 2018.

[19] RADFORD A, WU J, CHILD R, *et al*. Language models are unsupervised multitask learners[J]. *OpenAI Blog*, 2019, 1(8): 9.

[20] BROWN T, MANN B, RYDER N, *et al*. Language models are few-shot learners[C]. The 34th International Conference on Neural Information Processing Systems, Vancouver, Canada, 2020: 159.

[21] OpenAI. GPT-4 technical report[J]. arXiv preprint arXiv: 2303.08774, 2023.

[22] TOUVRON H, LAVRIL T, IZACARD G, *et al*. LLaMA: Open and efficient foundation language models[J]. arXiv preprint arXiv: 2302.13971, 2023.

[23] BAI Jinze, BAI Shuai, CHU Yunfei, *et al*. Qwen technical report[J]. arXiv preprint arXiv: 2309.16609, 2023.

[24] DeepSeek-AI. DeepSeek-V3 technical report[J]. arXiv preprint arXiv: 2412.19437, 2024.

[25] VASWANI A, SHAZEER N, PARMAR N, *et al*. Attention is all you need[C]. The 31st International Conference on Neural Information Processing Systems, Long Beach, USA, 2017: 6000–6010.

[26] LIU Haotian, LI Chunyuan, WU Qingyang, *et al*. Visual instruction tuning[C]. The 37th International Conference on Neural Information Processing Systems, New Orleans, USA, 2023: 1516.

[27] LIN Bin, YE Yang, ZHU Bin, *et al*. Video-LLaVA: Learning united visual representation by alignment before projection[C]. 2024 Conference on Empirical Methods in Natural Language Processing, Miami, USA, 2024: 5971–5984.

[28] KOH J Y, FRIED D, and SALAKHUTDINOV R R. Generating images with multimodal language models[C]. The 37th International Conference on Neural Information Processing Systems, New Orleans, USA, 2023: 939.

[29] ZHOU Yue, LAN Mengcheng, LI Xiang, *et al*. GeoGround: A unified large vision-language model for remote sensing visual grounding[J]. arXiv preprint arXiv: 2411.11904, 2024.

[30] WANG Peijin, HU Huiyang, TONG Boyuan, *et al*. RingMoGPT: A unified remote sensing foundation model for vision, language, and grounded tasks[J]. *IEEE Transactions on Geoscience and Remote Sensing*, 2025, 63: 5611320. doi: 10.1109/TGRS.2024.3510833.

[31] ZHAN Yang, XIONG Zhitong, and YUAN Yuan. SkyEyeGPT: Unifying remote sensing vision-language tasks via instruction tuning with large language model[J]. *ISPRS Journal of Photogrammetry and Remote Sensing*, 2025, 221: 64–77. doi: 10.1016/j.isprsjprs.2025.01.020.

[32] 张永军, 李彦胜, 党博, 等. 多模态遥感基础大模型: 研究现状与未来展望[J]. 测绘学报, 2024, 53(10): 1942–1954. doi: 10.11947/j.AGCS.2024.20240019.
ZHANG Yongjun, LI Yansheng, DANG Bo, *et al*. Multi-modal remote sensing large foundation models: Current research status and future prospect[J]. *Acta Geodaetica et Cartographica Sinica*, 2024, 53(10): 1942–1954. doi: 10.11947/j.AGCS.2024.20240019.

[33] HONG Danfeng, HAN Zhu, YAO Jing, *et al*. SpectralFormer: Rethinking hyperspectral image classification with transformers[J]. *IEEE Transactions on Geoscience and Remote Sensing*, 2022, 60: 5518615. doi: 10.1109/TGRS.2021.3130716.

[34] HONG Danfeng, ZHANG Bing, LI Xuyang, *et al*. SpectralGPT: Spectral remote sensing foundation model[J]. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024, 46(8): 5227–5244. doi: 10.1109/TPAMI.2024.3362475.

[35] FULLER A, MILLARD K, and GREEN J R. CROMA: Remote sensing representations with contrastive radar-optical masked autoencoders[C]. The 37th International Conference on Neural Information Processing Systems, New Orleans, USA, 2023: 241.

[36] WANG Yi, ALBRECHT C M, BRAHAM N A A, *et al*. Decoupling common and unique representations for multimodal self-supervised learning[C]. The 18th European Conference on Computer Vision, Milan, Italy, 2024: 286–303. doi: 10.1007/978-3-031-73397-0_17.

[37] 张良培, 张乐飞, 袁强强. 遥感大模型: 进展与前瞻[J]. 武汉大学学报(信息科学版), 2023, 48(10): 1574–1581. doi: 10.13203/j.whugis20230341.
ZHANG Liangpei, ZHANG Lefei, and YUAN Qiangqiang. Large remote sensing model: Progress and prospects[J]. *Geomatics and Information Science of Wuhan University*, 2023, 48(10): 1574–1581. doi: 10.13203/j.whugis20230341.

[38] CHIANG W L, LI Zhuohan, LIN Zi, *et al*. Vicuna: An open-source Chatbot impressing GPT-4 with 90%* ChatGPT quality[EB/OL]. https://lmsys.org/blog/2023-03-30-vicuna/, 2023.

[39] DeepSeek-AI. DeepSeek-R1: Incentivizing reasoning capability in LLMs via reinforcement learning[J]. arXiv preprint arXiv: 2501.12948, 2025.

[40] ALAYRAC J B, DONAHUE J, LUC P, *et al*. Flamingo: A visual language model for few-shot learning[C]. The 36th International Conference on Neural Information Processing Systems, New Orleans, USA, 2022: 1723.

[41] LI Junnan, LI Dongxu, SAVARESE S, *et al*. BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models[C]. The 40th International Conference on Machine Learning, Honolulu, USA, 2023: 814.

[42] KUCKREJA K, DANISH M S, NASEER M, *et al*. GeoChat: Grounded large vision-language model for remote sensing[C]. The IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, USA, 2024: 27831–27840. doi: 10.1109/CVPR52733.2024.02629.

[43] PANG Chao, WENG Xingxing, WU Jiang, *et al*. VHM: Versatile and honest vision language model for remote sensing image analysis[C]. The 39th AAAI Conference on Artificial Intelligence, Philadelphia, USA, 2025: 6381–6388. doi: 10.1609/aaai.v39i6.32683.

[44] LUO Junwei, PANG Zhen, ZHANG Yongjun, *et al*. SkySenseGPT: A fine-grained instruction tuning dataset and model for remote sensing vision-language understanding[J]. arXiv preprint arXiv: 2406.10100, 2024.

[45] SILVA J D, MAGALHÃES J, TUIA D, *et al*. Large language models for captioning and retrieving remote sensing images[J]. arXiv preprint arXiv: 2402.06475, 2024.

[46] GUO Mingning, WU Mengwei, SHEN Yuxiang, *et al*. IFShip: Interpretable fine-grained ship classification with domain knowledge-enhanced vision-language models[J]. *Pattern Recognition*, 2025, 166: 111672. doi: 10.1016/j.patcog.2025.111672.

[47] ZHANG Wei, CAI Miaoxin, ZHANG Tong, *et al*. Popeye: A unified visual-language model for multisource ship detection from remote sensing imagery[J]. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 2024, 17: 20050–20063. doi: 10.1109/JSTARS.2024.3488034.

[48] DENG Pei, ZHOU Wenqian, and WU Hanlin. ChangeChat: An interactive model for remote sensing change analysis via multimodal instruction tuning[C]. ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Hyderabad, India, 2025: 1–5. doi: 10.1109/ICASSP49660.2025.10890620.

[49] NOMAN M, AHSAN N, NASEER M, *et al*. CDChat: A large multimodal model for remote sensing change description[J]. arXiv preprint arXiv: 2409.16261, 2024.

[50] IRVIN J A, LIU E R, CHEN J C, *et al*. TEOChat: A large vision-language assistant for temporal earth observation data[C]. The 13th International Conference on Learning Representations, Singapore, Singapore, 2025.

[51] RADFORD A, KIM J W, HALLACY C, *et al*. Learning transferable visual models from natural language supervision[C]. The 38th International Conference on Machine Learning, 2021: 8748–8763.

[52] LIU Zhuang, MAO Hanzi, WU Chaoyuan, *et al*. A ConvNet for the 2020s[C]. The IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, USA, 2022: 11966–11976. doi: 10.1109/CVPR52688.2022.01167.

[53] SUN Quan, FANG Yuxin, WU L, *et al*. EVA-CLIP: Improved training techniques for CLIP at scale[J]. arXiv preprint arXiv: 2303.15389, 2023.

[54] DOSOVITSKIY A, BEYER L, KOLESNIKOV A, *et al*. An image is worth 16x16 words: Transformers for image recognition at scale[C]. The 9th International Conference on Learning Representations, 2021.

[55] LU Kaixuan, ZHANG Ruiqian, HUANG Xiao, *et al*. Aquila: A hierarchically aligned visual-language model for enhanced remote sensing image comprehension[J]. arXiv preprint arXiv: 2411.06074, 2024.

[56] LU Kaixuan. Aquila-plus: Prompt-driven visual-language models for pixel-level remote sensing image understanding[J]. arXiv preprint arXiv: 2411.06142, 2024.

[57] ZI Xing, NI Tengjun, FAN Xianjing, *et al*. AeroLite: Tag-guided lightweight generation of aerial image captions[J]. arXiv preprint arXiv: 2504.09528, 2025.

[58] JIANG Hongxiang, YIN Jihao, WANG Qixiong, *et al*. EagleVision: Object-level attribute multimodal LLM for remote sensing[J]. arXiv preprint arXiv: 2503.23330, 2025.

[59] SONI S, DUDHANE A, DEBARY H, *et al*. EarthDial: Turning multi-sensory earth observations to interactive dialogues[C]. The Computer Vision and Pattern Recognition Conference, Nashville, USA, 2025: 14303–14313.

[60] ZHANG Wei, CAI Miaoxin, NING Yaqian, *et al*. EarthGPT-X: Enabling MLLMs to flexibly and comprehensively understand multi-source remote sensing imagery[J]. arXiv preprint arXiv: 2504.12795, 2025.

[61] ZHANG Wei, CAI Miaoxin, ZHANG Tong, *et al*. EarthMarker: A visual prompting multimodal large language model for remote sensing[J]. *IEEE Transactions on Geoscience and Remote Sensing*, 2025, 63: 5604219. doi: 10.1109/TGRS.2024.3523505.

[62] WANG Fengxiang, CHEN Mingshuo, LI Yueying, *et al*. GeoLLaVA-8K: Scaling remote-sensing multimodal large language models to 8K resolution[J]. arXiv preprint arXiv: 2505.21375, 2025.

[63] LI Zhenshi, MUHTAR D, GU Feng, *et al*. LHRS-Bot-Nova: Improved multimodal large language model for remote sensing vision-language interpretation[J]. *ISPRS Journal of Photogrammetry and Remote Sensing*, 2025, 227: 539–550. doi: 10.1016/j.isprsjprs.2025.06.003.

[64] BAZI Y, BASHMAL L, AL RAHHAL M M, *et al*. RS-LLaVA: A large vision-language model for joint captioning and question answering in remote sensing imagery[J]. *Remote Sensing*, 2024, 16(9): 1477. doi: 10.3390/rs16091477.

[65] LIU Xu and LIAN Zhouhui. RSUniVLM: A unified vision

language model for remote sensing via granularity-oriented mixture of experts[J]. arXiv preprint arXiv: 2412.05679, 2024.

[66] KARANFIL E, IMAMOGLU N, ERDEM E, *et al*. A vision-language framework for multispectral scene representation using language-grounded features[J]. arXiv preprint arXiv: 2501.10144, 2025.

[67] ZHANG Hao, LI Feng, LIU Shilong, *et al*. DINO: DETR with improved denoising anchor boxes for end-to-end object detection[C]. The 11th International Conference on Learning Representations, Kigali, Rwanda, 2023.

[68] ZHAI Xiaohua, MUSTAFA B, KOLESNIKOV A, *et al*. Sigmoid loss for language image pre-training[C]. The IEEE/CVF International Conference on Computer Vision, Paris, France, 2023: 11941–11952. doi: 10.1109/ICCV51070.2023.01100.

[69] ZHANG Jihai, QU Xiaoye, ZHU Tong, *et al*. CLIP-MoE: Towards building mixture of experts for CLIP with diversified multiplet upcycling[J]. arXiv preprint arXiv: 2409.19291, 2024.

[70] WANG Weihan, LV Qingsong, YU Wenmeng, *et al*. CogVLM: Visual expert for pretrained language models[C]. The 38th International Conference on Neural Information Processing Systems, Vancouver, Canada, 2024: 3860.

[71] LIU Haotian, LI Chunyuan, LI Yuheng, *et al*. Improved baselines with visual instruction tuning[C]. The IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, USA, 2024: 26286–26296. doi: 10.1109/CVPR52733.2024.02484.

[72] KAUFMANN T, WENG P, BENGS V, *et al*. A survey of reinforcement learning from human feedback[J]. *Transactions on Machine Learning Research*, 2025, 2025.

[73] HU E J, SHEN Yelong, WALLIS P, *et al*. LoRA: Low-rank adaptation of large language models[C]. The 10th International Conference on Learning Representations, 2022.

[74] QU Bo, LI Xuelong, TAO Dacheng, *et al*. Deep semantic understanding of high resolution remote sensing image[C]. 2016 International Conference on Computer, Information and Telecommunication Systems (CITS), Kunming, China, 2016: 1–5. doi: 10.1109/CITS.2016.7546397.

[75] LU Xiaoqiang, WANG Binqiang, ZHENG Xiangtao, *et al*. Exploring models and data for remote sensing image caption generation[J]. *IEEE Transactions on Geoscience and Remote Sensing*, 2018, 56(4): 2183–2195. doi: 10.1109/TGRS.2017.2776321.

[76] LOBRY S, MARCOS D, MURRAY J, *et al*. RSVQA: Visual question answering for remote sensing data[J]. *IEEE Transactions on Geoscience and Remote Sensing*, 2020, 58(12): 8555–8566. doi: 10.1109/TGRS.2020.2988782.

[77] CHENG Qimin, HUANG Haiyan, XU Yuan, *et al*. NWPU-captions dataset and MLCA-Net for remote sensing image captioning[J]. *IEEE Transactions on Geoscience and Remote Sensing*, 2022, 60: 5629419. doi: 10.1109/TGRS.2022.3201474.

[78] YUAN Zhiqiang, ZHANG Wenkai, FU Kun, *et al*. Exploring a fine-grained multiscale method for cross-modal remote sensing image retrieval[J]. *IEEE Transactions on Geoscience and Remote Sensing*, 2022, 60: 4404119. doi: 10.1109/TGRS.2021.3078451.

[79] XIA Guisong, HU Jingwen, HU Fan, *et al*. AID: A benchmark data set for performance evaluation of aerial scene classification[J]. *IEEE Transactions on Geoscience and Remote Sensing*, 2017, 55(7): 3965–3981. doi: 10.1109/TGRS.2017.2685945.

[80] YANG Yi and NEWSAM S. Bag-of-visual-words and spatial extensions for land-use classification[C]. The 18th SIGSPATIAL International Conference on Advances in Geographic Information Systems, San Jose, USA, 2010: 270–279. doi: 10.1145/1869790.1869829.

[81] DAI Dengxin and YANG Wen. Satellite image classification via two-layer sparse coding with biased image representation[J]. *IEEE Geoscience and Remote Sensing Letters*, 2011, 8(1): 173–176. doi: 10.1109/LGRS.2010.2055033.

[82] SUMBUL G, CHARFUELAN M, DEMIR B, *et al*. Bigearthnet: A large-scale benchmark archive for remote sensing image understanding[C]. IGARSS 2019-2019 IEEE international geoscience and remote sensing symposium, Yokohama, Japan, 2019: 5901–5904. doi: 10.1109/IGARSS.2019.8900532.

[83] GUPTA R, GOODMAN B, PATEL N, *et al*. Creating xBD: A dataset for assessing building damage from satellite imagery[C]. The IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, Long Beach, USA, 2019: 10–17.

[84] CHRISTIE G, FENDLEY N, WILSON J, *et al*. Functional map of the world[C]. The IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, USA, 2018: 6172–6180. doi: 10.1109/CVPR.2018.00646.

[85] ZHANG Meimei, CHEN Fang, and LI Bin. Multistep question-driven visual question answering for remote sensing[J]. *IEEE Transactions on Geoscience and Remote Sensing*, 2023, 61: 4704912. doi: 10.1109/TGRS.2023.3312479.

[86] ZHAN Yang, XIONG Zhitong, and YUAN Yuan. RSVG: Exploring data and models for visual grounding on remote sensing data[J]. *IEEE Transactions on Geoscience and Remote Sensing*, 2023, 61: 5604513. doi: 10.1109/TGRS.2023.3250471.

[87] CHENG Gong, HAN Junwei, and LU Xiaoqiang. Remote sensing image scene classification: Benchmark and state of the art[J]. *Proceedings of the IEEE*, 2017, 105(10): 1865–1883. doi: 10.1109/JPROC.2017.2675998.

[88] LI Haifeng, JIANG Hao, GU Xin, *et al*. CLRS: Continual learning benchmark for remote sensing image scene classification[J]. *Sensors*, 2020, 20(4): 1226. doi: 10.3390/s20041226.

[89] ZHOU Zhuang, LI Shengyang, WU Wei, *et al*. NaSC-TG2: Natural scene classification with Tiangong-2 remotely sensed imagery[J]. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 2021, 14: 3228–3242. doi: 10.1109/JSTARS.2021.3063096.

[90] SUN Yuxi, FENG Shanshan, LI Xutao, *et al*. Visual grounding in remote sensing images[C]. The 30th ACM International Conference on Multimedia, Lisboa, Portugal, 2022: 404–412. doi: 10.1145/3503161.3548316.

[91] LI Xiang, DING Jian, and ELHOSEINY M. VRSBench: A versatile vision-language benchmark dataset for remote sensing image understanding[C]. The 38th International Conference on Neural Information Processing Systems, Vancouver, Canada, 2024: 106.

[92] ZHU Qiqi, ZHONG Yanfei, ZHAO Bei, *et al*. Bag-of-visual-words scene classifier with local and global features for high spatial resolution remote sensing imagery[J]. *IEEE Geoscience and Remote Sensing Letters*, 2016, 13(6): 747–751. doi: 10.1109/LGRS.2015.2513443.

[93] HELBER P, BISCHKE B, DENGEL A, *et al*. EuroSAT: A novel dataset and deep learning benchmark for land use and land cover classification[J]. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 2019, 12(7): 2217–2226. doi: 10.1109/JSTARS.2019.2918242.

[94] ZHU B, LUI N, IRVIN J, *et al*. METER-ML: A multi-sensor earth observation benchmark for automated methane source mapping[C]. The 2nd Workshop on Complex Data Challenges in Earth Observation (CDCEO 2022) Co-Located with 31st International Joint Conference on Artificial Intelligence and the 25th European Conference on Artificial Intelligence (IJCAI-ECAI 2022), Vienna, Austria, 2022: 33–43.

[95] LI Kun, VOSSELMAN G, and YANG M Y. HRVQA: A visual question answering benchmark for high-resolution aerial images[J]. *ISPRS Journal of Photogrammetry and Remote Sensing*, 2024, 214: 65–81. doi: 10.1016/j.isprsjprs.2024.06.002.

[96] HOXHA G and MELGANI F. A novel SVM-based decoder for remote sensing image captioning[J]. *IEEE Transactions on Geoscience and Remote Sensing*, 2022, 60: 5404514. doi: 10.1109/TGRS.2021.3105004.

[97] ZHENG Xiangtao, WANG Binqiang, DU Xingqian, *et al*. Mutual attention inception network for remote sensing visual question answering[J]. *IEEE Transactions on Geoscience and Remote Sensing*, 2022, 60: 5606514. doi: 10.1109/TGRS.2021.3079918.

[98] BASHMAL L, BAZI Y, AL RAHHAL M M, *et al*. CapERA: Captioning events in aerial videos[J]. *Remote Sensing*, 2023, 15(8): 2139. doi: 10.3390/rs15082139.

[99] SUN Xian, WANG Peijin, LU Wanxuan, *et al*. RingMo: A remote sensing foundation model with masked image modeling[J]. *IEEE Transactions on Geoscience and Remote Sensing*, 2023, 61: 5612822. doi: 10.1109/TGRS.2022.3194732.

[100] BI Hanbo, FENG Yingchao, TONG Boyuan, *et al*. RingMoE: Mixture-of-modality-experts multi-modal foundation models for universal remote sensing image interpretation[J]. arXiv preprint arXiv: 2504.03166, 2025.

[101] CHEN Hao and SHI Zhenwei. A spatial-temporal attention-based method and a new dataset for remote sensing image change detection[J]. *Remote Sensing*, 2020, 12(10): 1662. doi: 10.3390/rs12101662.

[102] SHI Qian, LIU Mengxi, LI Shengchen, *et al*. A deeply supervised attention metric-based network and an open aerial image dataset for remote sensing change detection[J]. *IEEE Transactions on Geoscience and Remote Sensing*, 2022, 60: 5604816. doi: 10.1109/TGRS.2021.3085870.

[103] LIU Chenyang, ZHAO Rui, CHEN Hao, *et al*. Remote sensing image change captioning with dual-branch transformers: A new method and a large scale dataset[J]. *IEEE Transactions on Geoscience and Remote Sensing*, 2022, 60: 5633520. doi: 10.1109/TGRS.2022.3218921.

[104] XIA Guisong, BAI Xiang, DING Jian, *et al*. DOTA: A large-scale dataset for object detection in aerial images[C]. The IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, USA, 2018: 3974–3983. doi: 10.1109/CVPR.2018.00418.

[105] LI Ke, WAN Gang, CHENG Gong, *et al*. Object detection in optical remote sensing images: A survey and a new benchmark[J]. *ISPRS Journal of Photogrammetry and Remote Sensing*, 2020, 159: 296–307. doi: 10.1016/j.isprsjprs.2019.11.023.

[106] ZHAO Yuanxin, ZHANG Mi, YANG Bingnan, *et al*. LuoJiaHOG: A hierarchy oriented geo-aware image caption dataset for remote sensing image–text retrieval[J]. *ISPRS Journal of Photogrammetry and Remote Sensing*, 2025, 222: 130–151. doi: 10.1016/j.isprsjprs.2025.02.009.

[107] YUAN Zhenghang, XIONG Zhitong, MOU Lichao, *et al*. ChatEarthNet: A global-scale image-text dataset empowering vision-language geo-foundation models[J]. *Earth System Science Data*, 2025, 17(3): 1245–1263. doi: 10.5194/essd-17-1245-2025.

[108] LI Haodong, ZHANG Xiaofeng, and QU Haicheng. DDFAV: Remote sensing large vision language models dataset and evaluation benchmark[J]. *Remote Sensing*, 2025, 17(4): 719. doi: 10.3390/rs17040719.

[109] AN Xiao, SUN Jiaxing, GUI Zihan, *et al*. COREval: A comprehensive and objective benchmark for evaluating the remote sensing capabilities of large vision-language models[J]. arXiv preprint arXiv: 2411.18145, 2024.

[110] ZHOU Yue, FENG Litong, LAN Mengcheng, *et al*. GeoMath: A benchmark for multimodal mathematical reasoning in remote sensing[C]. The 13th International Conference on Representation Learning, Singapore, Singapore, 2025.

[111] GE Junyao, ZHANG Xu, ZHENG Yang, *et al*. RSTeller: Scaling up visual language modeling in remote sensing with rich linguistic semantics from openly available data and large language models[J]. *ISPRS Journal of Photogrammetry and Remote Sensing*, 2025, 226: 146–163. doi: 10.1016/j.isprsjprs.2025.05.002.

[112] DU Siqi, TANG Shengjun, WANG Weixi, *et al*. Tree-GPT: Modular large language model expert system for forest remote sensing image understanding and interactive analysis[J]. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, 2023, XLVIII-1-W2-2023: 1729–1736. doi: 10.5194/isprs-archives-XLVIII-1-W2-2023-1729-2023.

[113] GUO Haonan, SU Xin, WU Chen, *et al*. Remote sensing ChatGPT: Solving remote sensing tasks with ChatGPT and visual models[C]. IGARSS 2024-2024 IEEE International Geoscience and Remote Sensing Symposium, Athens, Greece, 2024: 11474–11478. doi: 10.1109/IGARSS53475.2024.10640736.

[114] LIU Chenyang, CHEN Keyan, ZHANG Haotian, *et al*. Change-agent: Towards interactive comprehensive remote sensing change interpretation and analysis[J]. *IEEE Transactions on Geoscience and Remote Sensing*, 2024, 62: 5635616. doi: 10.1109/TGRS.2024.3425815.

[115] XU Wenjia, YU Zijian, MU Boyang, *et al*. RS-Agent: Automating remote sensing tasks through intelligent agents[J]. arXiv preprint arXiv: 2406.07089, 2024.

[116] SINGH S, FORE M, and STAMOULIS D. GeoLLM-Engine: A realistic environment for building geospatial copilots[C]. The IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, Seattle, USA, 2024: 585–594. doi: 10.1109/CVPRW63382.2024.00063.

[117] SINGH S, FORE M, and STAMOULIS D. Evaluating tool-augmented agents in remote sensing platforms[J]. arXiv preprint arXiv: 2405.00709, 2024.

作 者 简 介

许文嘉，博士，副教授，主要研究方向为通信遥感一体化与遥感智能解译。

于睿卿，博士，主要研究方向为遥感多模态大语言模型。

薛铭浩，硕士，主要研究方向为遥感多模态大语言模型。

张源奔，博士，副研究员，主要研究方向为时空数据决策智能、数字地球与数字孪生。

魏智威，博士，讲师，主要研究方向为地理可视化、LLM4Urban、时空图谱、三维重建等。

张　柘，博士，研究员，主要研究方向为新体制SAR成像与信号处理技术。

彭木根，博士，教授，主要研究方向为无线移动通信和低轨信息通信网络。

(责任编辑：于青)