中图法分类号:TP751　　**文献标识码**: A　　**文章编号**: 1006-8961(XXXX)XX-0001-20

论文引用格式:ZHANG Jingnan, ZHANG Fengjun. Retrieval-augmented and parameter-efficient fine-tuning framework for multi-task semantic understanding of remote sensing imagery[J/OL]. Journal of Image and Graphics, XXXX:1-20. DOI: 10.11834/jig.260289.(张婧楠, 张逢骏. 融合检索增强与参数高效微调的遥感影像多任务语义理解框架[J/OL]. 中国图象图形学报,XXXX:1-20. DOI: 10.11834/jig.260289.)[DOI:10.11834/jig.260289]

融合检索增强与参数高效微调的遥感影像多任务语义理解框架

张婧楠,张逢骏

江苏科技大学计算机学院,江苏省镇江市　212100

摘　要:**目的**　针对高分辨率遥感影像在通用多模态大模型中存在视觉token序列过长、细粒度地物信息易丢失、遥感领域知识表达不足等问题,本文构建一种基于检索增强与参数高效微调的遥感影像多任务语义理解框架FM9G4BV-RS。**方法**　该框架以FM9G4BV为基座模型,首先通过自适应高分辨率切片策略将大幅面遥感影像划分为结构保持性较好的图像块序列,并结合Grouped-Query Attention(GQA)提高长视觉序列下的跨模态对齐效率;其次,采用Low-Rank Adaptation(LoRA)对遥感任务相关投影层进行参数高效微调,并利用NF4量化降低模型存储与推理开销;最后,引入Retrieval-Augmented Generation(RAG)机制,将遥感文献、数据集说明和灾害报告等外部知识作为可追溯上下文融入语言生成过程,从而增强模型在专业遥感场景中的语义表达能力。**结果**　在VRSBench视觉语言评测基准上的实验结果表明,FM9G4BV-RS在图像描述、视觉定位和遥感视觉问答任务上均优于基座模型。其中,BLEU-1、BLEU-2和BLEU-4分别达到34.8%、24.2%和10.1%,较基座模型分别提升18.4、16.6和8.4个百分点;METEOR和ROUGE-L分别达到35.6和37.2,较基座模型分别提高10.3和19.4;Acc@0.5(IoU All)达到30.7%,VQA(All)达到47.3%。进一步的组件级消融结果表明,自适应切片、LoRA微调和RAG知识增强对视觉定位、语义生成和问答准确性具有主要贡献,GQA与NF4量化则主要提升长序列推理效率和资源适配能力。**结论**　本文方法能够在不全量重训练基础模型的条件下提升遥感多模态语义理解性能,为大尺寸遥感影像的图文生成、视觉定位、交互式问答和多时相变化理解等任务提供了一种可扩展的轻量化适配方案。

关键词:多模态大模型;遥感影像语义理解;检索增强生成(RAG);参数高效微调(LoRA);视觉语言模型;多任务适配

Retrieval-augmented and parameter-efficient fine-tuning framework for multi-task semantic understanding of remote sensing imagery

ZHANG Jingnan, ZHANG Fengjun

School of Computer Science, Jiangsu University of Science and Technology, Zhenjiang City, Jiangsu Province, 212100, China

Abstract: **Objective** With the rapid development of Earth observation technologies, remote sensing imagery has become one of the most important data sources for large-scale environmental monitoring, urban planning, disaster assessment, military reconnaissance, agricultural management, and ecological analysis. Recent advances in multimodal large language models and vision-language foundation models have provided a new paradigm for remote sensing image interpretation by

收稿日期:2026-05-21;**修回日期**:2026-08-05

基金项目:国家自然科学基金(项目编号:62404085)

Supported by: Project supported by the National Natural Science Foundation of China (Grant No. 62404085)

enabling unified modeling of visual and linguistic information. However, directly applying general-purpose multimodal large models to remote sensing imagery still faces several critical limitations. First, high-resolution remote sensing images usually contain extremely large spatial dimensions and dense visual details, which produce excessively long visual-token sequences and lead to high computational overhead during inference. Second, fine-grained semantic structures such as small targets, narrow roads, airports, ships, and buildings are easily lost during conventional image resizing or patch encoding processes. Third, general multimodal models often lack professional remote sensing knowledge and therefore exhibit insufficient semantic reasoning ability in domain-specific tasks such as disaster analysis, multi-temporal change understanding, and visual grounding. In addition, full-parameter fine-tuning of large models requires massive computational resources, which limits practical deployment on medium- and low-resource hardware platforms. To address these issues, this study proposes FM9G4BV-RS, a retrieval-augmented and parameter-efficient fine-tuning framework for multi-task semantic understanding of remote sensing imagery. **Method** The proposed framework is developed on top of the FM9G4BV multimodal foundation model and integrates adaptive high-resolution image slicing, Grouped-Query Attention (GQA), Retrieval-Augmented Generation (RAG), Low-Rank Adaptation (LoRA), and NF4 quantization into a unified remote sensing multimodal learning architecture. First, to overcome the computational bottleneck caused by ultra-high-resolution remote sensing images, an adaptive image slicing strategy is introduced. Instead of directly resizing large images into fixed resolutions, the framework dynamically partitions input images into structure-preserving image blocks according to image scale, aspect ratio, and spatial semantic distribution. A layout scoring mechanism considering area coverage, information balance, and boundary preservation is designed to select the optimal slicing configuration, thereby reducing visual-token redundancy while preserving critical spatial structures. Subsequently, the sliced image patches are encoded into high-dimensional visual tokens through the visual encoder. To improve cross-modal interaction efficiency under long visual-token sequences, the framework introduces Grouped-Query Attention into the multimodal Transformer architecture. Compared with conventional Multi-Head Attention, GQA reduces computational redundancy by allowing multiple query heads to share grouped key-value representations, thus improving memory efficiency and inference speed while maintaining semantic alignment capability. In addition, the framework adopts LoRA-based parameter-efficient fine-tuning for remote sensing task adaptation. Instead of updating all model parameters, LoRA inserts trainable low-rank matrices into the query, key, and value projection layers of the Transformer, significantly reducing the number of trainable parameters and lowering training costs. To further improve deployment efficiency on resource-constrained hardware platforms, NF4 quantization and double quantization strategies are incorporated during inference. These methods compress model weights into low-bit representations while preserving the statistical distribution characteristics of neural network parameters, thereby reducing GPU memory consumption and storage overhead. Furthermore, Retrieval-Augmented Generation (RAG) is integrated by encoding remote sensing knowledge sources into a vector database, enabling task-relevant knowledge retrieval to enhance domain-specific semantic reasoning. During inference, the model retrieves task-relevant knowledge fragments according to semantic similarity and integrates them into the generation process as traceable contextual evidence. This mechanism effectively alleviates hallucination problems and improves semantic consistency in professional remote sensing scenarios. The proposed framework supports a wide range of remote sensing vision-language tasks, including image captioning, visual grounding, visual question answering, object detection, disaster assessment, semantic scene understanding, and multi-temporal change interpretation. Experiments are conducted using multiple remote sensing datasets, including Git-10M, RSICD, VRSBench, FAIR1M, NWPU VHR-10, LEVIR-CC, CC-Foundation, OpenEarthMap-SAR, BRIGHT, and xBD. Among them, VRSBench is selected as the primary benchmark for quantitative evaluation because it simultaneously supports image captioning, visual grounding, and visual question answering tasks. **Result** Experimental results demonstrate that FM9G4BV-RS achieves significant performance improvements over the original FM9G4BV foundation model and several representative multimodal baselines in remote sensing semantic understanding tasks. On the VRS-Bench benchmark, the proposed framework achieves BLEU-1, BLEU-2, and BLEU-4 scores of 34.8%, 24.2%, and 10.1%, respectively, improving the base model by 18.4, 16.6, and 8.4 percentage points. The METEOR and ROUGE-L scores reach 35.6 and 37.2, respectively, indicating substantial improvements in semantic consistency, sentence fluency, and structural coherence. In the visual grounding task, the framework achieves an Acc@0.5 (IoU All) score of 30.7%,

which significantly exceeds the base model and demonstrates enhanced spatial localization capability under complex remote sensing scenes. For remote sensing visual question answering, the VQA (All) accuracy reaches 47.3%, showing improved multimodal reasoning ability. Qualitative analysis further confirms the effectiveness of the proposed framework. In airport and harbor scenes containing dense small targets, FM9G4BV-RS accurately identifies airplanes and ships while generating more precise localization boundaries than the baseline models. In disaster assessment scenarios, the framework successfully combines visual evidence and external knowledge to generate semantically consistent descriptions of damaged regions and disaster severity. In multi-temporal change understanding tasks, the framework can capture semantic differences between temporal image pairs and produce natural language explanations for urban expansion, building changes, and land-use transitions. Ablation studies demonstrate the contribution of each core module. Removing LoRA causes the largest performance degradation, indicating that parameter-efficient domain adaptation is essential for remote sensing semantic learning. Removing the RAG module reduces METEOR and ROUGE-L scores, suggesting that external knowledge retrieval effectively enhances professional semantic generation and factual consistency. Replacing adaptive slicing with fixed slicing reduces visual grounding performance because fixed layouts are less effective at preserving spatially important targets. Replacing GQA with MQA decreases both localization and language generation performance, demonstrating that grouped key-value sharing achieves a better balance between efficiency and representation capability. Although FP16 inference slightly outperforms NF4 quantization in accuracy, NF4 significantly reduces memory requirements and improves deployment flexibility on edge devices. **Conclusion** This study proposes FM9G4BV-RS, a retrieval-augmented and parameter-efficient fine-tuning framework for multi-task semantic understanding of remote sensing imagery. By combining adaptive high-resolution image slicing, GQA-based cross-modal interaction, LoRA parameter-efficient fine-tuning, NF4 quantization, and RAG-based external knowledge enhancement, the proposed framework effectively improves semantic reasoning capability, visual-language alignment performance, and deployment efficiency in remote sensing scenarios without requiring full retraining of the foundation model. Experimental results on multiple remote sensing benchmarks demonstrate that FM9G4BV-RS achieves superior performance in image captioning, visual grounding, visual question answering, and multi-temporal change understanding tasks. Compared with conventional multimodal large models, the proposed framework provides a scalable and lightweight adaptation solution for remote sensing semantic understanding under limited computational resources. The framework also exhibits strong extensibility toward future remote sensing applications such as disaster emergency response, interactive geospatial intelligence analysis, cross-modal retrieval, and autonomous Earth observation systems. Nevertheless, several limitations remain. Some tasks, including disaster assessment and fine-grained object classification, still require more comprehensive quantitative evaluation on larger standardized datasets. In addition, the framework may still encounter localization deviations and semantic ambiguity in extremely dense small-target scenes and complex background environments. Future work will focus on expanding multimodal remote sensing knowledge bases, improving temporal-spatial alignment mechanisms, introducing uncertainty-aware reasoning methods, and exploring real-time deployment strategies under edge-cloud collaborative environments.

Key words: multimodal large models; semantic understanding of remote sensing imagery; Retrieval-Augmented Generation (RAG); parameter-efficient fine-tuning (LoRA); vision-language model; multi-task adaptation

论文引用格式:DOI:10.11834/jig.260289

0 引　言

遥感影像作为获取地表空间信息的重要数据源,在城市规划、生态监测、灾害评估和国土资源管理等领域发挥着关键作用(Wang等,2025)。伴随卫星与航空遥感技术的快速发展,高分辨率、多光谱、合成孔径雷达(Synthetic Aperture Radar,SAR)以及激光雷达(LiDAR)等多源异构数据不断增长,这些数据在空间细节与语义丰富性上均呈现显著提升,同时也带来了更大的计算复杂性和分析难度(Wang等,2025)。大量遥感影像数据的积累对自动化信息提取、深层语义理解提出了更高要求,传统基于人工设计特征的解译方法难以满足这种大规模、多样性场景下的需求。

近年来,随着深度学习技术在遥感影像处理中的广泛应用,卷积神经网络与视觉 Transformer 等架

构在目标检测、地物分类与变化检测等任务中取得了显著进展，但是这些方法大多针对特定任务进行设计，模型泛化能力和跨任务协同能力仍然有限（Zhou等，2025；Zheng等，2024）。相比之下，多模态大模型通过统一视觉信息与语言语义的框架，为通用遥感图像理解提供了新的技术路线（Wei等，2025）。已有研究表明，多模态模型在高分辨率遥感影像的视觉问答、图像描述生成等任务中可以显著改善语义表达能力，并推动不同任务间的知识迁移（Dang等，2025；Ge等，2025；Tao等，2025）。例如，在遥感特定场景下，通过动态分辨率输入策略和多尺度视觉-语言对齐机制，可在保持关键细节的同时提升跨模态一致性，这对复杂地物解释尤为重要（Zhang等，2025）。

尽管上述多模态方法在遥感智能解译方面展现了潜力，但它们仍面临几个核心挑战（Wang等，2024）：一是遥感影像具有极高的空间分辨率，将完整图像直接输入模型不仅对算力提出巨大压力，也可能导致细粒度空间信息的损失；二是遥感场景中语义理解往往依赖专业领域知识，而通用视觉-语言模型在这类知识表达方面能力不足；三是高效推理在不同算力平台上的适配仍是工程实现的难点。因此，如何设计既能高效处理大尺寸图像、又能融合专门领域知识的多模态模型，是当前遥感智能解译研究必须解决的问题。

针对上述问题，本文构建了一种面向遥感影像分析与语义理解的多模态大模型框架FM9G4BV-RS。该框架以视觉语言模型为基础，通过自适应高分辨率图像切片策略实现大尺寸遥感影像的高效编码，并利用Grouped-Query Attention机制增强视觉特征与语言特征之间的交互能力。类似研究也提出了通过联合训练多任务数据来提升模型泛化能力的方法（Li等，2026）。同时，在模型推理过程中引入检索增强生成（Retrieval-Augmented Generation，RAG）机制，将遥感领域知识库信息融入语义生成过程，以改善专业场景下的表达质量和准确性。此外，为了降低模型训练和部署开销，本文结合LoRA参数高效微调与NF4量化技术对模型进行轻量化优化，使得模型在不同算力平台上实现高效训练与推理。

本文的主要贡献如下：1）针对大幅高分辨率遥感影像直接输入多模态大模型时计算开销大、细节信息易损失的问题，提出自适应切片与GQA协同的跨模态编码策略，在控制视觉token规模的同时保留关键地物结构；2）针对遥感知识具有传感器差异、空间尺度变化和时空关联等特点，本文构建了包含传感器类型、空间分辨率、区域与时间信息、任务类型及地物类别等元数据的遥感知识库，并采用“向量召回—元数据过滤—任务约束重排序”的检索流程。筛选后的知识片段作为可追溯的领域上下文，与视觉token和任务指令共同输入多模态语言模型；3）将所提框架应用于图像描述、视觉定位、视觉问答和多时相变化理解等遥感视觉语言任务，并在VRS-Bench基准上验证其有效性，同时明确该框架可向灾害场景问答和交互式遥感解译等任务扩展。

1 模型架构与核心模块

1.1 总体架构

FM9G4BV-RS是一款面向遥感多模态智能理解任务构建的统一大模型框架，支持遥感影像与自然语言指令的端到端协同建模与推理。模型整体架构由图像预处理模块、视觉编码模块、多模态语言生成模块以及跨模态交互机制构成，其整体架构如图1所示。

在输入层，系统支持多源异构遥感数据，包括合成孔径雷达（SAR）影像、多光谱影像、红外影像、RGB影像、激光雷达数据及全色影像等多种模态，具备良好的跨传感器兼容能力。不同模态数据首先进入图像预处理模块，在该阶段完成图像尺寸标准化、分辨率统一及切片处理等操作，以消除数据分布差异对模型建模的影响。经过预处理后的影像数据被输入视觉编码器进行特征提取。该模块基于深度视觉模型，将原始像素信息映射为高维语义表示（visual tokens）。随后，在跨模态建模阶段，这些视觉token与文本输入（如任务指令或问题描述）共同嵌入统一特征空间，并输入基于Transformer架构构建的多模态语言模型。模型通过跨模态注意力机制实现视觉信息与语义信息的深度融合，从而生成符合任务需求的输出结果。在统一建模框架下，视觉链路负责完成多源遥感影像的归一化、自适应切片、视觉token编码和GQA跨模态融合；知识链路负责从遥感知识库中检索与当前任务相关的文本证据，并通过元数据过滤和重排序形成检索上下文。两条链路在多模态语言模型的统一上下文窗口内汇合，

使模型能够同时利用细粒度图像证据和遥感领域知识完成生成与定位任务。

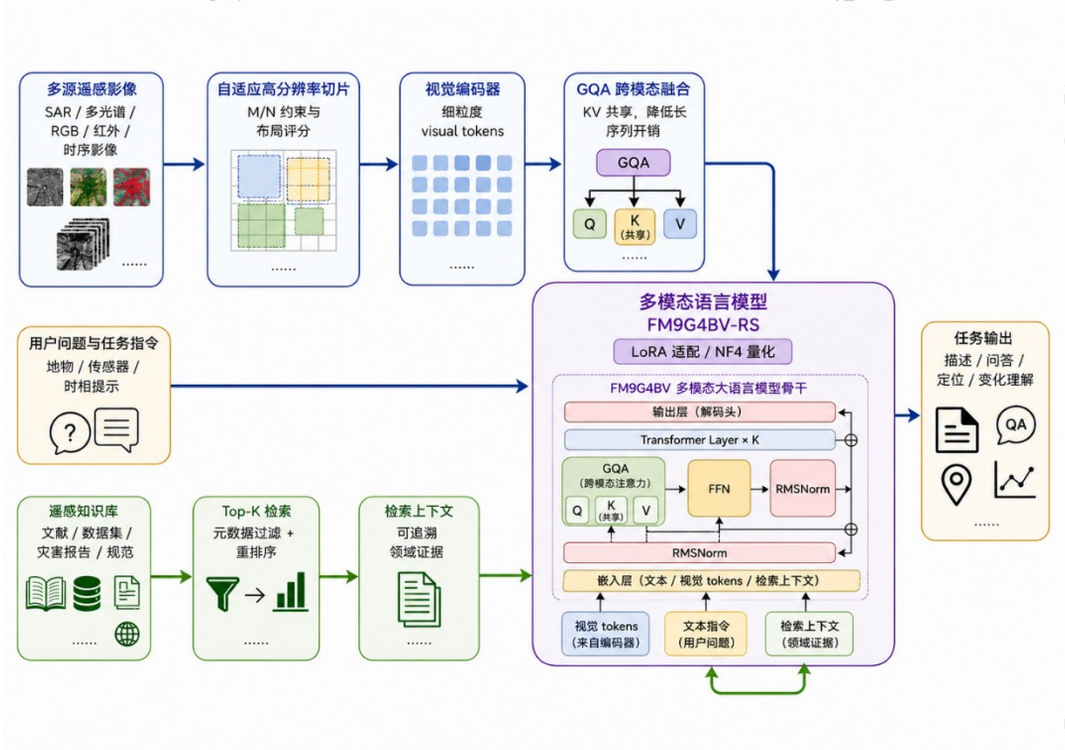


图 1　FM9G4BV-RS 模型整体架构流程图

Fig 1　FM9G4BV-RS Model Overall Architecture Flowchart

为进一步展示模型在实际遥感任务中的推理能力，图 2 给出了 FM9G4BV-RS 在典型遥感任务中的应用示例，直观展示了 FM9G4BV-RS 如何对多源遥感图像进行语义解析与精确生成，覆盖了视觉问答、视觉定位、图文对话、时序变化检测、环境语义分析等多个任务类别。模型接收文本输入形式的查询，如“该区域是否存在车辆”等，可基于输入遥感图像输出坐标、类别等信息，部分结果含边界框及多轮推理结果，体现其较强的泛化能力与多任务处理一致性。图 3 的内容展示了 FM9G4BV-RS 在真实遥感应用场景下的语言交互能力、目标识别精度以及对多源数据（如 SAR/红外）的高质量适配能力，为其在灾害应急、生态监测、军事侦察等领域的落地应用提供了技术支撑与验证。

1.2　高分辨率遥感图像建模

遥感影像通常具有分辨率高、覆盖范围大的特点，单幅影像尺寸可达到数万像素。若直接缩放后输入视觉模型，不仅会增加显存和计算开销，还可能造成飞机、车辆和船舶等小目标的细节丢失。为此，本文在 FM9G4BV-RS 中引入自适应高分辨率图像切片策略，根据模型可处理的视觉 token 数量和输入影像尺寸，动态确定切片数量与网格布局。

根据输入图像尺寸、模型最大切片数 M 以及视觉编码器输入分辨率估计理想切片数 N；随后枚举满足 $n\_h \times n\_w \leqslant M$ 的候选网格布局 $\boldsymbol{G}=\{n\_h, n\_w\}$，如 $1\times N$、2×3、3×2 等，并结合图像长宽比进行初筛。为避免固定切片造成的目标截断和信息不均衡，本文定义布局评分函数 $S(\boldsymbol{G})=\lambda 1 S\_\mathrm{area}(\boldsymbol{G})+\lambda 2 S\_\mathrm{balance}(\boldsymbol{G})+\lambda 3 S\_\mathrm{boundary}(\boldsymbol{G})$，其中 S_area 衡量切片后有效区域覆盖程度，S_balance 衡量各切片信息量均衡性，S_boundary 用于惩罚重要目标或显著边缘被切断的情况。文本输入和系统提示分别占用 T_{text} 和 T_{sys} 个 token，单个图像切片产生 T_{slice} 个视觉 token，则最大切片数需满足：$[M \leq \left\lfloor \frac{L_{\max}-T_{\mathrm{text}}-T_{\mathrm{sys}}}{T_{\mathrm{slice}}} \right\rfloor$，结合 KV 缓存、batch size 和设备显存限制，以保证模型能够稳定运行。理想切片数（N）根据原始图像尺寸和视觉编码器的标准输入分辨率估计。设原始图像尺寸为

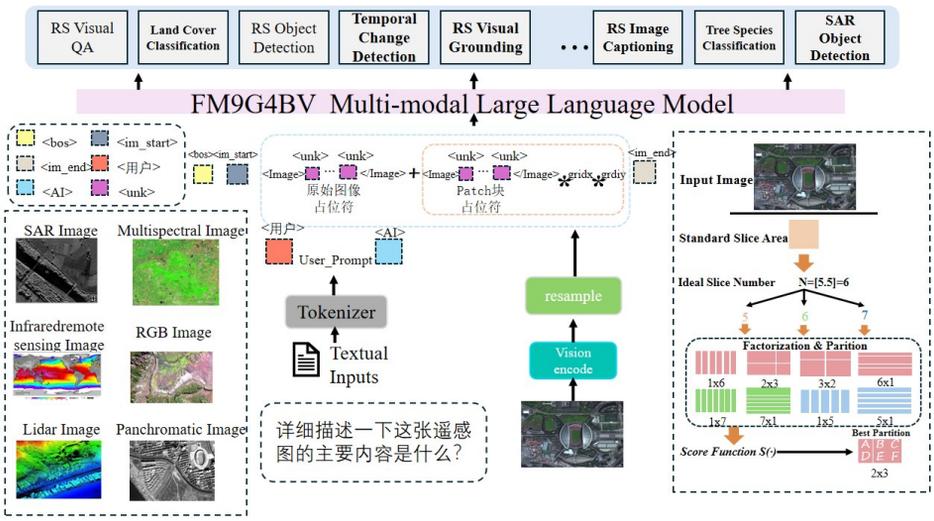


图 2　技术总览

Fig 2　Technology Overview

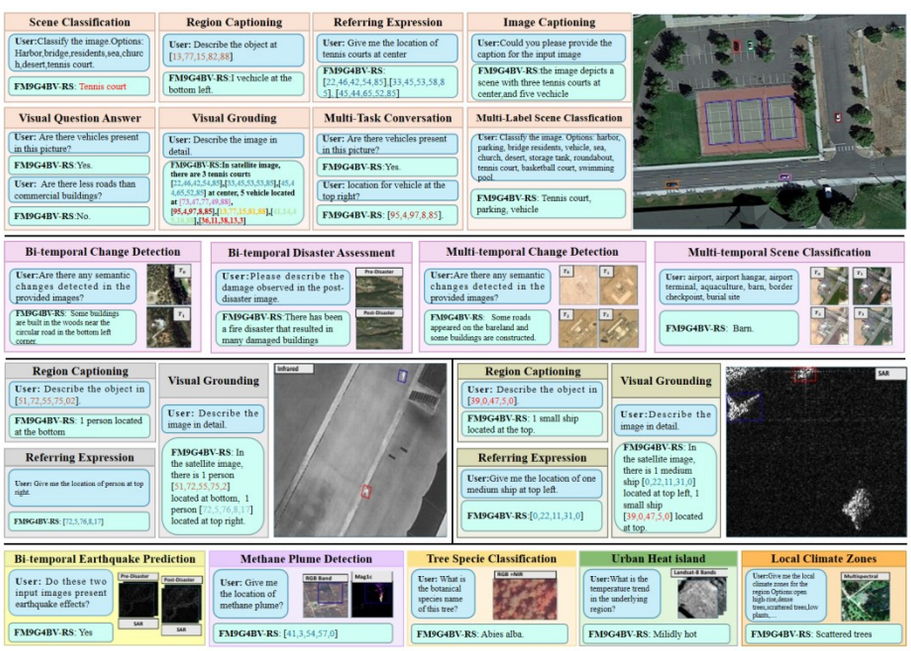


图 3　多模态遥感大模型 FM9G4BV-RS 在典型任务中的推理效果

Fig 3　Inference Performance of the Multimodal Remote Sensing Large Model FM9G4BV-RS in Typical Tasks

$H\times W$，视觉编码器输入尺寸为$R\times R$，则初始切片数为：$[N_0=\left\lceil\frac{H\times W}{R^2}\right\rceil$。$N=\min(M,\max(1,N_0))$，枚举满足$n_h\times n_w\leq M$的候选切片网格，并综合考虑图像长宽比、有效区域覆盖率、视觉 token 开销和目标跨边界风险，选择得分最高的布局作为最终切片方案。实验中$\lambda1$、$\lambda2$、$\lambda3$分别设置为0.4、0.3和0.3，并选择评分最高的候选布局作为最终切片方案。能够在控制计算量和显存开销的同时，尽可能保留高分辨率遥感影像中的细粒度地物信息。随后，各切片通过 ReSample 模块统一分辨率，并输入视觉编码器生成高维视觉 token。该方法能够在控制视觉 token 数量、计算开销和显存占用的同时，尽可能保留高分辨率遥感影像中的细粒度地物信息。

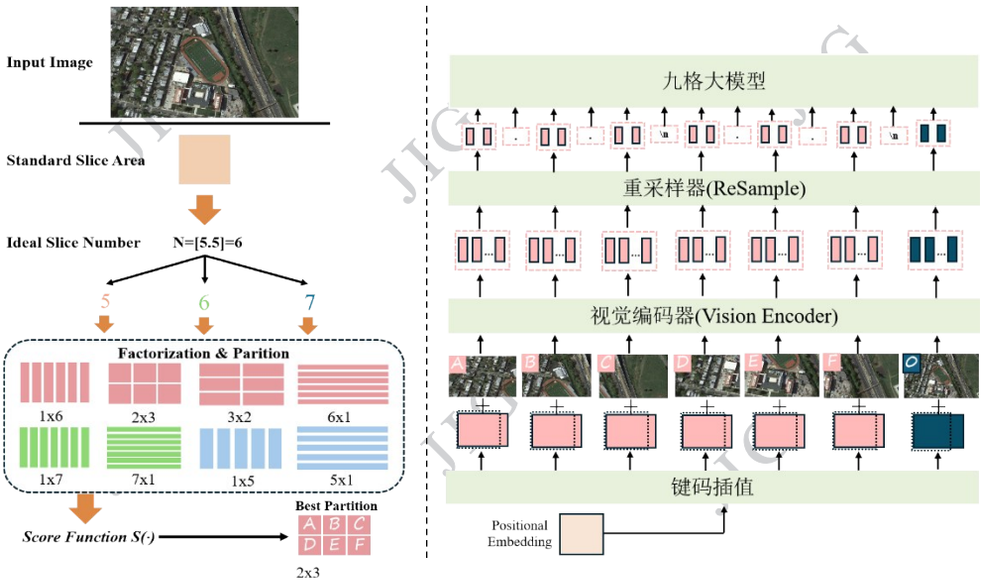


图 4　遥感图像的切片与编码流程结构图

Fig 4　Flowchart of the Tiling and Encoding Process for Remote Sensing Images

1.3　跨模态特征融合机制

在多模态大模型中，视觉信息与语言信息的有效融合是实现复杂语义理解的关键。FM9G4BV-RS采用基于Transformer的跨模态建模结构，引入了Grouped-Query Attention（GQA）机制作为其Transformer层的关键对齐模块，以有效提升在大规模遥感图像切片与语言任务中的跨模态推理效率。

传统多头注意力机制（Multi-Head Attention）中，每个查询需与独立的键值对计算注意力，存在显著的计算冗余。而GQA通过将多个查询头划分为若干组并共享键值向量，在保持局部感知能力的同时大幅压缩计算开销，提升了在长序列条件下的显存效率与运行时性能。

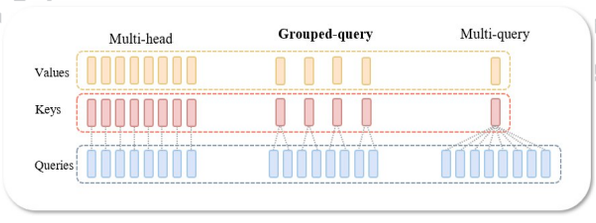


图 5　三种注意力机制（多头、分组、多查询）的结构对比

Fig 5　A Structural Comparison of Three Attention Mechanisms (Multi-Head, Grouped, and Multi-Query)

在模型执行流程中，高分辨率图像经切片后生成的细粒度视觉token与文本嵌入token一同送入跨模态Transformer层的GQA模块。该模块通过分组共享机制强化了图文对齐效率，提升了对目标的语义聚焦与空间定位能力。跨模态特征融合与注意力机制在遥感多模态任务中被证明能够显著提高图文对齐精度和空间语义理解（Zhang等，2024）。图5展示了Grouped-query Attention（GQA）机制相较于Multi-head Attention（MHA）和Multi-query Attention（MQA）在键值共享结构上的差异。GQA将所有查询头划分为 G 个分组，每个分组共享一组键（Key）和值（Value）向量，从而在注意力表达能力与计算效率之间取得平衡。当 G=1（即GQA-1）时，所有查询头共享同一组键值，退化为MQA；当 G=H（其中 H 为查询头总数）时，每个查询头对应独立的键值对，等价于标准MHA。因此，GQA可以视为在MHA与MQA之间的中间形式，提供灵活的配置空间来调节模型的性能与资源消耗。

此外，跨模态Transformer层中的前馈网络采用SiLU激活函数，进一步提升模型在复杂语义推理任务中的表达能力。

1.4　检索增强生成技术（RAG）

遥感影像解译不仅依赖影像中的光谱、纹理和空间结构信息，还涉及传感器成像特性、空间分辨率、地物类别定义、区域背景、成像时间及灾害判读标准等领域知识。通用多模态大模型主要依赖预训练阶段形成的参数化知识，在处理专业术语、特定传感器或区域性任务时，容易出现知识不足、概念混淆

和事实偏差。为此，本文引入检索增强生成(Retrieval-Augmented Generation，RAG)机制，为模型补充与当前影像和任务相关的外部知识。在知识库构建阶段，首先对遥感文献、数据集说明、地物类别体系、灾害报告和技术规范等资料进行清洗与分段，并为各知识片段保留标题、来源、传感器类型、空间分辨率、区域与时间、任务类型及地物类别等元数据。随后，利用文本编码器将知识片段映射为向量表示并存入知识库。对于用户问题或任务提示，系统生成查询向量，先基于语义相似度召回 Top-K 候选片段，再结合传感器类型、任务类别、区域时间信息和关键词约束进行过滤与重排序，从而获得与当前遥感场景更匹配的知识内容。与通用向量检索相比，该流程进一步考虑了遥感数据的传感器差异、空间尺度和时空属性，可减少语义相关但成像条件或任务背景不一致的知识片段进入模型上下文。在生成阶段，遥感影像经视觉编码器转换为视觉 token，检索片段则被组织为带有来源标识的文本 token。视觉 token、用户问题 token 和检索文本 token 共同输入多模态语言模型，并在统一上下文窗口内通过自注意力或 GQA 机制完成信息交互。检索内容不直接替代模型的视觉判断，而是作为领域证据对视觉信息进行补充和约束，从而提高图像描述、遥感问答、灾害解释和变化理解等任务的专业性、可靠性和可解释性。系统采用离线更新与在线检索相结合的方式：离线阶段定期补充遥感资料，在线阶段根据影像和任务提示实时检索相关知识。遥感 RAG 与视觉 token 的交互过程如图6所示。

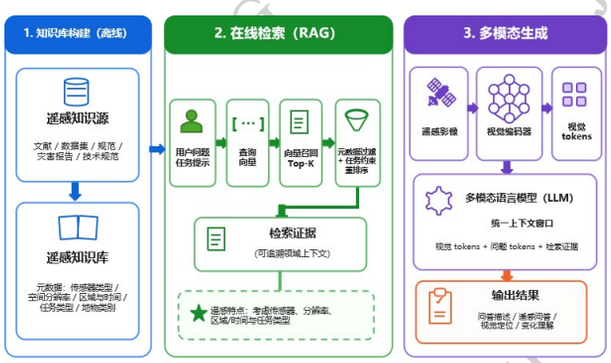


图 6　遥感RAG与视觉交互流程图

Fig 6　Remote sensing RAG and visual interaction flowchart

2　遥感任务适配与模型优化

2.1　多任务遥感建模

2.1.1　图像目标检测

目标检测是遥感影像智能解译中的基础任务，其目标是在图像中识别感兴趣目标并确定其类别与空间位置，为地理实体提取、变化监测及语义分析提供重要支撑。然而，在遥感场景中该任务仍面临多种挑战，例如大尺寸影像的分块处理、小目标与密集目标识别困难以及检测精度与计算效率之间的平衡问题。

为提升复杂场景下的检测能力，本文在 FM9G4BV-RS 模型中引入视觉—语言协同建模机制，实现多模态条件下的目标检测。系统支持遥感影像与文本提示的联合输入，其中图像模态提供空间结构与纹理信息，文本模态提供目标类别或语义约束，从而形成跨模态语义引导。如图 7 所示，遥感图像首先输入视觉编码器以提取高维视觉特征，随后与文本提示共同输入多模态语言模型进行联合建模，最终输出目标类别及对应边界框坐标。

在输出阶段，模型采用结构化文本形式表示检测结果。每个目标实例通过类别标签及对应边界框坐标进行描述，从而兼顾自然语言表达能力与结构化数据的可解析性。该方式便于后续自动化标注、结果评估及可视化处理，可支持大规模遥感影像的目标检测任务。

模型结构方面，视觉编码器首先提取多尺度视觉特征，并将其映射为统一的语义表示；随后该表示与文本特征共同输入大语言模型，通过跨模态注意力机制实现视觉与语言信息的深度融合。最终，模型输出目标类别与空间位置，并可进一步扩展至图像问答或图像描述等相关任务。

2.1.2　多时相遥感图像变化检测

多时相遥感图像变化检测旨在识别同一区域在不同时间获取的图像之间的差异，从而揭示地表变化特征及其演化过程。该技术在土地利用监测、城市扩张分析及灾害评估等领域具有重要应用价值。然而，该任务仍面临多方面挑战，包括多相时影像的精准配准、多源遥感数据融合以及光照变化、季节差异等非结构性因素带来的干扰。此外，变化检测不仅需要识别变化区域，还需进一步判别变化类型及

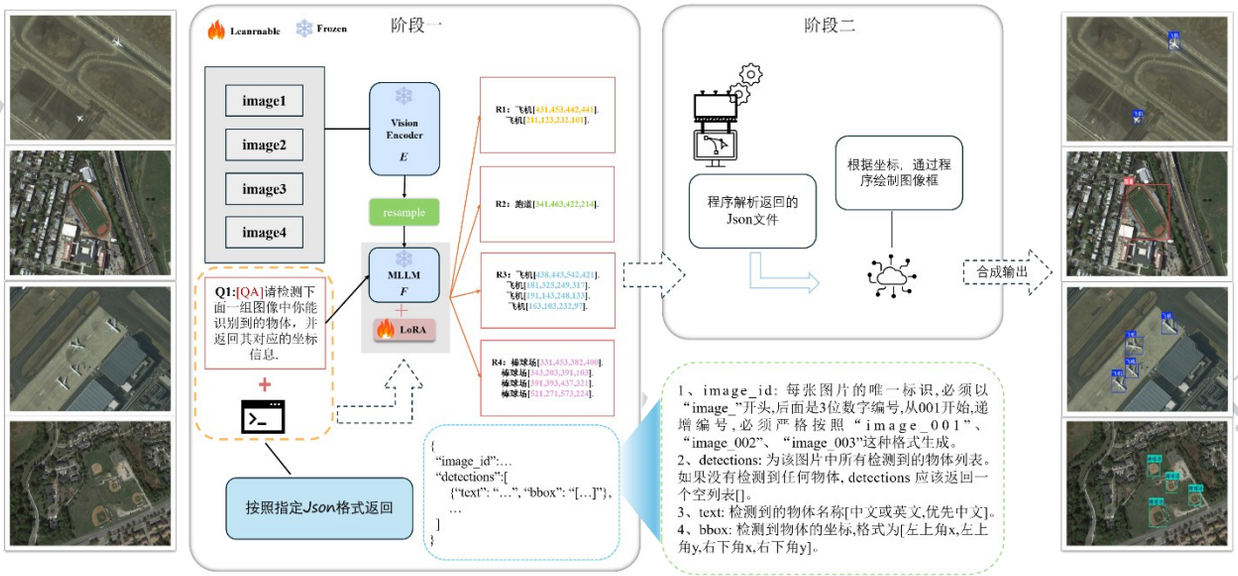


图 7　目标检测

Fig 7　Object Detection

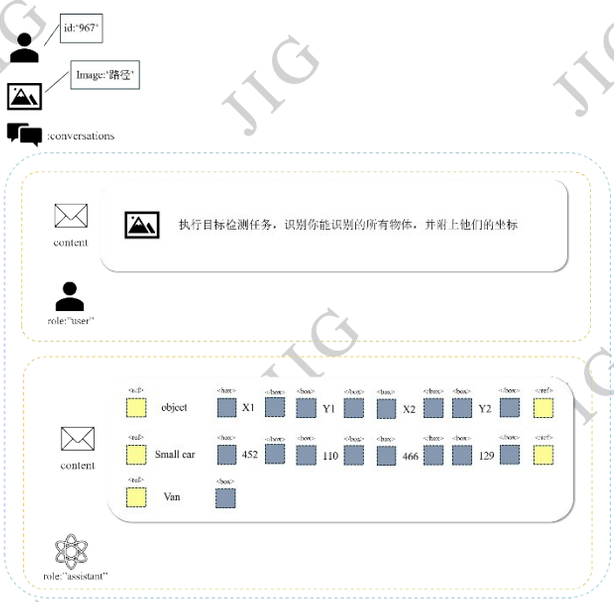


图 8　目标检测训练的数据集格式

Fig 8　Data set formats for object detection training

其语义属性。

图 9 中展示了一种面向多时相遥感图像理解任务的大模型推理架构。该模型以同一区域不同时刻获取的遥感影像序列为输入，通过共享视觉编码器分别提取各时相图像的视觉 token 表示，并在时间维度上引入轻量级时序 Transformer 进行跨帧建模。对于双时相输入 $\boldsymbol{I}\_t1$ 和 $\boldsymbol{I}\_t2$，模型首先获得对应视觉表示 z_t1 和 z_t2，随后构造差分特征 $\Delta z=z\_t2-z\_t1$，并将其与原始时相特征共同输入多模态语言模型。为降低配准误差、季节变化和光照差异带来的伪变化影响，本文在语义层面对双时相 token 进行对齐，使模型优先关注建筑物增减、道路变化、水体扩张等具有明确遥感语义的变化区域。在统一框架下，模型既可输出结构化变化结果，也可生成自然语言变化描述，实现"时序感知—语义对齐—语言生成"的闭环推理，适用于多时相遥感变化理解、土地利用演化分析及交互式问答等场景。

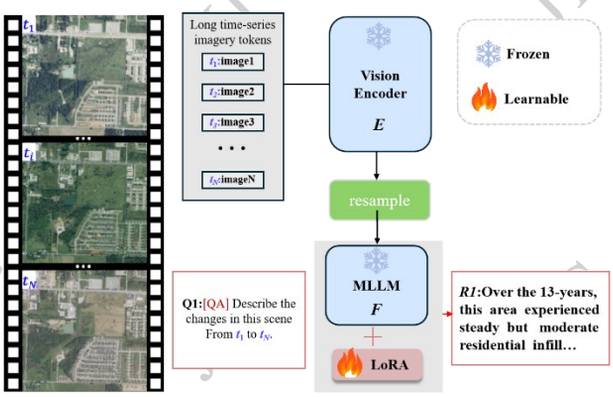


图 9　基于时序遥感图像的多模态语言模型时空推理

Fig 9　Multimodal Language Model-Based Spatiotemporal Inference Using Temporal Remote Sensing Images

实验数据集采用 CC-Foundation 与 LEVIR-CC。CC-Foundation 是规模较大的遥感变化描述数据集，包含约 20 万对遥感图像及 120 万条变化描述文本，覆盖多种地貌及长时间尺度变化场景，为模型训练提供了图像与语言之间的语义对齐数据。LEVIR-

CC主要关注建筑变化检测任务,包含10077对遥感图像及5万余条变化描述,记录同一区域在不同时间的空间演化过程,并对变化位置与类型进行标注,常用于变化检测与图文生成任务评估(Liu等,2022)。

2.1.3 地物分类与灾害评估

地物分类是遥感影像解译的重要任务,其目标是根据影像的空间结构与光谱信息识别不同类型的地表覆盖对象。在复杂地表环境中,地物类别之间往往存在较高的相似性,同时还受到尺度变化和边界模糊等因素影响,从而增加了分类难度。针对上述问题,FM9G4BV-RS通过融合遥感影像特征与文本语义信息构建多模态分类框架。模型在视觉编码器中提取影像的空间结构与光谱特征,并结合文本描述及地理先验信息进行跨模态特征对齐,通过多模态注意力机制实现联合建模。该方法能够在复杂场景中区分相似地物类别,并支持地物识别、语义分割及多尺度目标分类等相关任务,为土地利用调查、生态环境监测及城市规划等应用提供技术支持。

在灾害评估任务中,多模态信息融合能够提升对灾情变化的识别能力。FM9G4BV-RS利用遥感影像、时序变化数据以及灾情文本信息进行联合分析,对地震、洪涝和火灾等灾害引起的地表变化进行识别,并对受灾区域进行语义划分与灾情等级分析。通过多时相影像特征与文本语义的协同建模,模型能够对灾害影响区域进行综合判读,为灾后评估与应急决策提供辅助信息。

实验中选取OpenEarthMap-SAR与BRIGHT数据集分别用于地物分类与灾害评估任务。OpenEarthMap-SAR是面向合成孔径雷达(SAR)遥感影像的地物分类数据集,包含35个区域和8类主要地物,并提供对应语义标注。SAR影像具有全天候成像能力,可作为光学遥感的重要补充,为模型提供多源信息以提高复杂环境下的分类性能(Chen等,2025)。

BRIGHT数据集主要用于遥感灾害评估研究,包含约4200对灾前—灾后遥感影像(光学与SAR)以及建筑损毁标注信息,覆盖地震、洪涝和火灾等多类灾害事件。该数据集通过结合灾前光学影像与灾后SAR影像,支持跨模态变化分析,可用于建筑损毁检测、灾情识别及语义分类等任务的模型训练与评估。

2.2 多模态模型训练与推理机制

为提升多模态大模型在遥感场景中的语义理解与推理能力,本文构建了面向遥感任务的统一训练与推理框架,通过多模态数据组织、参数高效微调以及多图与时序推理机制,实现模型在复杂遥感场景下的稳定学习与泛化应用。该框架不仅支持单图语义理解,还能够扩展至多图与视频序列推理,为多时相遥感分析与复杂场景认知提供技术基础。

2.2.1 单图多轮对话训练机制

在单图任务训练中,模型以单张遥感图像及其对应文本描述或对话作为输入,通过结构化数据格式实现图像与语言信息的统一组织。每条样本以字典形式存储,其中包含唯一图像路径及对应的多轮对话序列,使模型能够在固定视觉上下文下进行连续语义推理。对话内容按照用户与模型交替的方式组织,从而模拟真实的人机交互过程,并强化模型在图像理解基础上的语言生成能力。

在训练过程中,遥感图像首先经视觉编码器转换为高维视觉token表示,并通过特征重采样模块映射至语言模型可接受的嵌入空间。随后,这些视觉token与文本token共同输入多模态语言模型进行联合建模,使模型能够在单一视觉场景下完成目标识别、场景理解及自然语言生成等任务。该训练方式在保证图文语义对齐的同时,简化了跨模态特征融合过程,提高了模型训练效率与语义表达能力。

2.2.2 多图及视频推理机制

相比单图任务,多图推理需要同时处理多张图像并建模其之间的语义关联,因此对模型的跨图信息整合能力提出更高要求。在多图推理过程中,系统首先将多张图像与任务文本共同输入视觉编码器,对每张图像分别进行patch划分与特征提取,生成对应的高维视觉token表示。随后,通过特征重采样模块将视觉特征映射至统一语义空间,并引入跨图特征交互机制,使模型能够在不同图像之间建立上下文联系,从而捕捉对象对应关系、场景差异及潜在语义变化。

在特征融合阶段,多图视觉token与包含图像位置信息的文本token按序列形式输入多模态语言模型,通过条件语言建模实现跨图语义推理,使模型能够完成多图问答、图像对比分析及多视角场景理解等任务。该机制显著增强了模型在复杂遥感场景中的综合认知能力,尤其适用于多区域对比分析与多

源遥感信息融合场景。

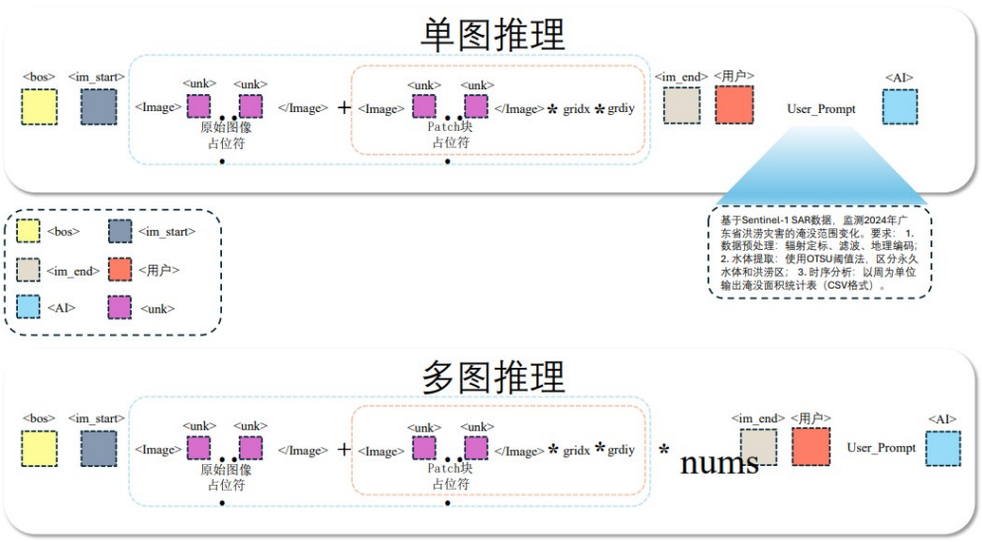


图 10　单图与多图输入推理示意图

Fig 10　Schematic Diagram of Single-Image and Multi-Image Input Inference

进一步地，视频推理可视为时序多图推理的扩展形式，其核心在于建模连续帧之间的动态变化关系。如图 11 所示，在视频处理过程中，首先对视频序列进行关键帧提取或均匀采样，并利用视觉编码器生成每帧的高维特征表示，从而构建时序视觉特征序列。随后，将包含时序任务指令的文本 token 与帧特征在时间维度上进行对齐，并输入多模态语言模型，通过跨帧注意力机制学习帧间依赖关系与动态变化模式，使模型能够完成事件过程分析、目标运动识别以及视频问答等时序推理任务。

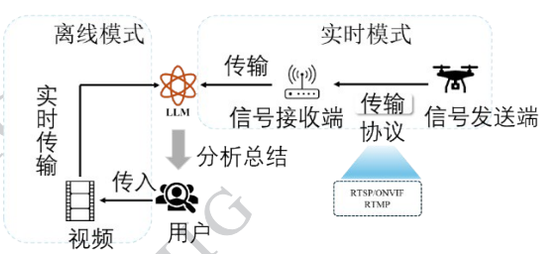


图 11　融合视觉语言建模的大模型视频目标检测流程图

Fig 11　Flowchart of a Large-Scale Model for Video Object Detection Incorporating Visual Language Modeling

2.2.3　多模态训练数据构建

本文选用 Git-10M 与 RSICD 两类遥感图文数据集构建训练数据，以支撑模型在遥感场景下的多模态语义学习能力。Git-10M 作为大规模遥感图文数据集，用于跨模态预训练，增强模型的开放域语义表达能力；RSICD 数据集作为遥感图像描述数据集，用于监督微调，强化模型对遥感图像内容与自然语言之间的语义对齐能力。

通过结合大规模开放域数据与高质量任务数据进行联合训练，模型能够在保持语义表达能力的同时，提高在复杂遥感场景中的理解与推理性能。

2.3　模型训练优化与高效部署策略

针对多模态大模型在遥感影像分析与语义理解任务中的适应能力提升问题，本文采用参数高效微调与量化推理优化相结合的训练策略。整体训练流程主要包括 LoRA 参数高效微调与模型量化与推理优化两个阶段，通过减少可训练参数规模并压缩模型存储开销，在保证模型性能的前提下提高训练与部署效率。

2.3.1　LoRA 参数高效微调

为降低多模态大模型微调过程中的显存占用与计算开销，本文采用参数高效微调（Parameter-Efficient Fine-Tuning，PEFT）方法，并选用 LoRA（Low-Rank Adaptation）实现模型适配。LoRA 通过冻结预训练模型参数，仅在 Transformer 层引入可训练的低秩矩阵，以低秩更新方式实现模型参数调整，在保留预训练知识的同时显著降低训练成本。

设预训练权重矩阵 $\boldsymbol{W} \in \mathbf{R}^{d\times k}$，LORA 引入可训练低秩矩阵对：

$$\Delta \boldsymbol{W} = \boldsymbol{B}\cdot\boldsymbol{A}\#(1)$$

其中$\boldsymbol{B}\in\mathbf{R}^{d\times r},\boldsymbol{A}\in\mathbf{R}^{r\times k},r\ll d$.

前向传播过程重构为：

$$\boldsymbol{h}=(\boldsymbol{W}+\alpha\cdot\boldsymbol{BA})\boldsymbol{x}\#(2)$$

式中α为缩放因子，用于平衡新知识与原始知识权重。

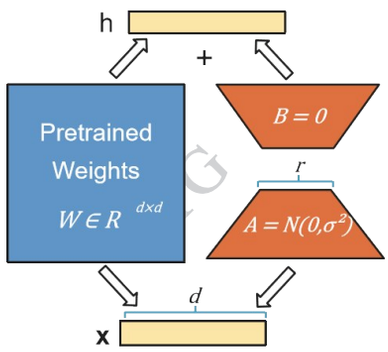


图 12　LoRA微调结构原理示意图

Fig 12　Schematic Diagram of the LoRA Fine-Tuning Architecture

通过将高维权重更新表示为低秩矩阵乘积，并根据任务需求调整秩参数r，可以在较低计算成本下实现模型能力的有效提升。与传统全量微调相比，该方法仅需训练少量新增参数，从而显著降低显存占用和训练开销。

2.3.2　NF4量化与推理优化

在完成LoRA微调后，为进一步降低模型部署成本并提升推理效率，本文在推理阶段引入模型量化技术。轻量化大模型技术已成为遥感领域的研究热点，尤其在地表异常探测等实时性要求高的任务中具有重要应用（Yan等，2025）。本文采用QLoRA框架中的4-bit NF4（NormalFloat Quantization）权重量化与双重量化（Double Quantization）策略，在保证模型性能的同时降低模型存储开销。

NF4针对神经网络权重近似服从正态分布的特点，采用非均匀分位量化以减小低比特量化误差。每个量化bin的中心值q_i计算公式为：

$$q_i=\frac{1}{2}\left(\mathrm{Q}_x\left(\frac{i}{2^k+1}\right)+\mathrm{Q}_x\left(\frac{i+1}{2^k+1}\right)\right)\#(3)$$

其中$k=4$表示量化位宽，$i\in\{0,1,\cdots,15\}$，$\mathrm{Q}_x(\cdot)$为标准正态分布的分位函数。

在NF4量化过程中，本文进一步引入“双重量化（Double Quantization）”方法，对量化过程中使用的scale参数进行再次量化，以降低量化常数的存储开销。

整体反量化过程可用以下公式描述：

$$\mathrm{doubleDequant}\left(c_1^{FP32},c_2^{k-bit},\boldsymbol{W}^{k-bit}\right)=\mathrm{dequant}\left(\mathrm{dequant}\left(c_2^{k-bit},c_1^{FP32}\right),\boldsymbol{W}^{k-bit}\right)\#(4)$$

其中，dequant表示基础反量化操作，满足：

一级反量化：$\mathrm{dequant}\left(c_2^{k-bit},c_1^{FP32}\right)=c_1^{FP32}\cdot c_2^{k-bit}=c_2^{FP32}$

二级反量化：$\mathrm{dequant}\left(\boldsymbol{W}^{k-bit},c_1^{FP32}\right)=c_2^{FP32}\cdot\boldsymbol{W}^{k-bit}$

最终得到高精度恢复权重（如BF16格式）。

在QLoRA框架中，采用以下配置：

权重$\boldsymbol{W}$使用4-bit NF4表示；

一级量化常数c_2使用FP8精度（块大小为256）；

二级量化常数c_1保持为FP32（块大小为64）；

所有反量化操作最终输出BF16精度张量。

在实际前向传播中，双重量化与LoRA模块共同参与推理，公式如下：

$$\boldsymbol{Y}^{BF16}=\boldsymbol{X}^{BF16}\cdot\mathrm{doubleDequant}\left(c_1^{FP32},c_2^{FP8},\boldsymbol{W}^{NF4}\right)+\boldsymbol{X}^{BF16}\cdot\boldsymbol{L}_1^{BF16}\cdot\boldsymbol{L}_2^{BF16}\#(5)$$

其中，$\boldsymbol{X}$为输入特征张量，$\boldsymbol{L}_1,\boldsymbol{L}_2$为LoRA的低秩可训练矩阵。

在训练过程中，仅通过$\boldsymbol{L}_1,\boldsymbol{L}_2$路径传播梯度，冻结主干权重$\boldsymbol{W}^{NF4}$及其所有量化参数，确保高效微调并减少显存占用。

3　实验设计和评估

3.1　数据集

本实验采用多源异构遥感数据集构建训练与评估体系，数据集选择如下：

3.2　评价指标

为系统评估FM9G4BV-RS在遥感多模态任务中的性能，本文采用多种通用评价指标，从文本生成质量、语义一致性及视觉区域定位精度等方面进行综合评估。对于文本生成任务，采用BLEU（1/2/4-gram）、METEOR和ROUGE-L评价生成文本与参考文本在词汇匹配、语义一致性及结构完整性等方面的表现；对于多项选择分类任务，采用准确率（Accuracy）评价模型预测结果与真实标签的一致性；对于视觉定位任务，采用交并比（Intersection over Union，IoU）衡量预测区域与真实区域的空间重叠程度。多指标联合评估能够从分类准确性、文本生成质量及

表 1　数据集用途、任务与评价指标对应关系

Table 1　Correspondence among Dataset Usage, Tasks and Evaluation Metrics

数据集	用途	任务	评价指标/输出
Git-10M	训练数据构建	遥感图文语义学习	图像描述文本
RSICD	训练与辅助评估	图像描述	BLEU、METEOR、ROUGE-L
VRSBench	主要定量测试	图像描述、视觉定位、VQA	BLEU、Acc@0.5、VQA、METEOR、ROUGE-L
FAIR1M	任务适配与案例展示	视觉定位/目标接地	边界框和IoU类指标
NWPU VHR-10	案例展示	典型地物目标识别	类别与位置
CC-Foundation	变化理解训练/扩展支撑	多时相变化描述	变化描述文本
LEVIR-CC	变化理解训练/案例展示	建筑变化理解	区域变化说明
OpenEarthMap-SAR	任务扩展说明	SAR地物分类	输出类别和语义标注
BRIGHT/xBD	灾害场景案例与知识库支撑	灾害评估问答	灾害等级、损毁描述与问答结果

空间定位精度等多个维度客观反映模型性能。

3.3　实施细节与训练策略

在多模态遥感大模型FM9G4BV-RS的训练过程中，综合考虑遥感影像跨模态语义对齐与生成任务对计算资源的需求，在保证模型性能的同时兼顾训练效率。训练与验证阶段均采用FP16半精度计算，以有效降低显存占用并提升计算吞吐效率。

在模型优化策略方面，针对遥感影像具有高分辨率与复杂语义结构的特点，对视觉编码器进行可训练微调，以增强模型对空间纹理与目标结构的表征能力；同时保持大语言模型主体参数冻结，仅在跨模态注意力层的Query、Key与Value投影模块中引入轻量级参数高效适配结构，从而在显著减少可训练参数规模的同时，实现视觉语义向语言空间的有效映射与融合。

优化器采用AdamW，并结合余弦退火学习率调度机制，以提升训练稳定性与收敛效率；训练过程中启用梯度检查点与DeepSpeed ZeRO-2显存优化策略，以降低显存占用并提高训练效率。

3.4　硬件设置

本文构建了涵盖高性能GPU服务器与边缘端设备的异构实验环境，从而评估FM9G4BV-RS在不同算力条件下的训练与推理性能，验证模型在多场景部署中的适应能力与可迁移性。实验所使用的主要硬件配置如表2所示。

表 2　实验硬件设置

Table 2　Experimental Hardware Setup

操作系统	GPU型号	GPU显存
Ubuntu 24.04.2 LTS(x86_64)	NVIDIA GeForce RTX 5090 D	32607MiB(32GB)
Windows11 22631.5472	NVIDIA GeForce RTX 3060	8192MiB(8GB)
Ubuntu 24.04.2 LTS(x86_64)	NVIDIA GeForce RTX 4070ti	12288 MiB (12GB)
MacBook Pro 2021	M1 Max MPS	UMA(16GB)

其中，高端GPU服务器主要用于模型训练与大规模推理评估，而中端显卡与移动端设备用于验证模型在资源受限场景下的推理效率与部署可行性。通过跨平台实验设置，进一步证明了FM9G4BV-RS在多算力环境中的稳定性与实际应用潜力。

4　模型训练

4.1　训练数据构建与管理

为满足多模态模型在不同遥感任务中的训练需

求，本文构建了统一的训练数据组织与管理方式。训练数据覆盖图像描述、视觉问答、目标检测、多时相变化检测及灾害评估等典型遥感任务，均采用JSON格式存储，通过统一字段关联图像与文本信息，实现多任务数据的统一管理和模型输入格式的一致性。

针对图像理解任务，本文采用统一的视觉—语言对话组织形式，其中单图任务完成单幅遥感影像的语义理解与问答，多图任务支持多幅遥感影像联合输入，实现跨视角、多场景信息的综合分析。

目标检测任务采用结构化标注方式，在文本响应中同时包含目标类别及对应边界框坐标，使模型能够联合完成目标识别与视觉定位。

变化检测与灾害评估任务均采用多时相遥感影像作为输入，并结合统一文本提示引导模型完成变化识别、目标定位及语义描述。其中，变化检测侧重于不同时相地物变化分析，灾害评估进一步融合灾前、灾后影像完成灾情识别与评估。

4.2 训练参数设置

本研究基于PyTorch深度学习框架对多模态遥感基础模型FM9G4B-V-RS进行有监督微调训练，训练过程中采用LoRA参数高效微调方法。

在训练过程中，视觉编码器部分允许参与微调以增强遥感影像特征表达能力，而大语言模型主体保持冻结，仅在自注意力层的Query、Key与Value投影模块中插入LoRA适配器，从而实现视觉语义特征向语言空间的有效映射。

此外，训练阶段采用FP16半精度计算以降低显存占用，并结合梯度累积与DeepSpeed ZeRO-2显存优化策略，使模型能够在中等算力条件下完成稳定训练。模型训练的主要参数配置如表3所示。

表3 训练参数设置

Table 3 Training Parameter Settings

类别	项目	配置说明
--model_max_length	4096	输入序列最大长度
--max_slice_nums	9	模型切片最大次数
--max_steps	10000	最大训练步数
--per_device_train_batch_size	2	每台设备训练batch size
--gradient_accumulation_steps	8	梯度累积步数，减少显存占用
--learning_rate	1e-6	学习率
--weight_decay	0.1	权重衰减系数
--warmup_ratio	0.01	学习率预热比例
--lr_scheduler_type	"cosine"	学习率调度策略
LoRA	enabled	参数高效微调
tune_vision_encoder	yes	微调视觉编码器
tune_llm	no	冻结语言模型
--gradient_checkpointing	true	启用梯度检查点以节省显存
--deepspeed	ds_config_zero2.json	使用DeepSpeed ZeRO-2配置

5 实验结果与分析

5.1 实验结果

本研究提出的FM9G4BV-RS模型在遥感基准数据集VRSBench的模型评估结果如表4所示，评估效果图如图13所示。与当前主流多模态视觉-语言模型的性能进行了系统性对比评估，覆盖BLEU-1、BLEU-2、BLEU-4、METEOR与ROUGE-L五个文本生成与语义一致性指标。结果表明，经过遥感领域任务微调后，FM9G4BV-RS在所有评估维度上均实现了对基座模型FM9G4BV的显著性能提升。其中，

BLEU-1、BLEU-2 与 BLEU-4 分别由 16.4%、7.6%、1.7% 提升至 34.8%、24.2%、10.1%，尤其 BLEU-4 提升 8.4 个百分点，显示其在捕捉细粒度信息与复杂句法结构方面的增强能力；METEOR 得分由 25.3 显著提高至 35.6，较 GPT-4V（20.9）高出 14.7 个百分点，优于 LlaVA-1.5（21.9）与 Mini-Gemini（21.5）；ROUGE-L 提升至 37.2，较基座模型（17.8）提升 19.4 分，并接近 LlaVA-1.5（36.9）与 Mini-Gemini（36.8）等高性能模型。整体来看，FM9G4BV-RS 在文本生成的准确性、长距离语义连贯性、同义表达泛化能力及空间关系描述完整性等方面均表现优越，同时保持了较低的推理计算开销，验证了其在遥感语义理解与跨模态对齐任务中的有效性与实用价值。

为保证对比实验的公平性，表 4 中的各模型均采用统一的 VRSBench 测试集、Prompt 模板、评价脚本及推理参数进行评估；闭源模型通过官方接口在相同测试样本和提示词条件下完成推理，其实验结果仅供对比参考。因此，表 4 主要用于比较不同模型在遥感图像描述、视觉定位和视觉问答任务上的相对表现，而不等同于标准目标检测、地物分类或灾害评估专项实验。

表 4　评估结果

Table 4　Evaluation Results

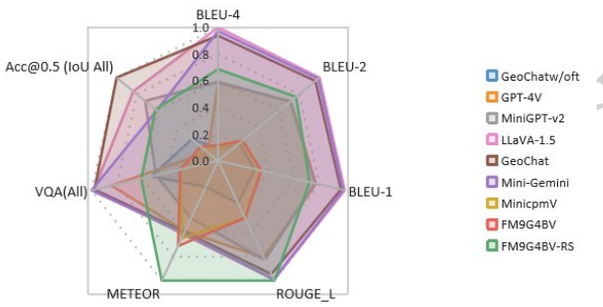
Model	BLEU-1	BLEU-2	BLEU-4	Acc@0.5 (IoU All)	VQA(All)	METEOR	ROUGE_L
GeoChatw/oft	13.9	6.6	1.4	12.9	40.8	7.8	13.2
GPT-4V	37.2	22.5	8.6	5.1	65.6	20.9	30.1
MiniGPT-v2	36.8	22.4	8.7	35.8	38.2	17.1	30.8
LlaVA-1.5	48.1	31.5	14.7	41.6	76.4	21.9	36.9
GeoChat	46.7	30.2	13.8	49.8	76.0	21.1	35.2
Mini-Gemini	47.6	31.1	14.3	30.1	77.8	21.5	36.8
MinicpmV	16.4	7.3	1.5	8.3	23.1	23.0	17.7
FM9G4BV	16.4	7.6	1.7	8.4	23.3	25.3	17.8
FM9G4BV-RS	34.8	24.2	10.1	30.7	47.3	35.6	37.2

在定性评估方面，图 14 与图 15 直观展示了 FM9G4BV-RS 在遥感影像多模态视觉接地（Visual Grounding）任务中的优越性。在图 14 所示的机场场景中，FM9G4BV-RS 能够精准捕捉跑道及停机坪上的飞机目标，其生成的预测框与地面真值（Ground Truth）高度重合，显著优于出现明显漏检的基座模型 FM9G4BV 及 MiniCPMV，且在定位边界的紧凑性上优于 GPT-4V 与 LlaVa-1.5。而在图 15 的近海船舶检测任务中，面对海面复杂背景纹理与目标尺度差异，FM9G4BV-RS 依然保持了极高的鲁棒性，能够准确识别并定位不同方位的船只，有效避免了如 GeoChat 或 Mini-Gemini 在类似场景下出现的检测框偏移或比例失真问题。这一系列可视化结果有力地印证了表 4 的量化数据，证明了 FM9G4BV-RS 在经过遥感领域针对性微调后，不仅在文本生成指标上表现卓越，在处理地物目标的视觉感知与空间对齐任务时也具备更强的工程实用价值。

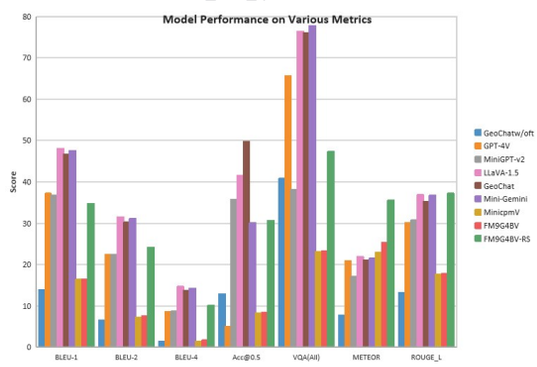
5.2　比较分析

为进一步系统性评估 FM9G4BV-RS 在遥感场景下自然语言生成、图像语义理解与空间定位等多任务中的综合性能，本研究选取 GPT-4V、LlaVA-1.5、MiniGPT-v2、Mini-Gemini、GeoChat、MinicpmV 以及基座模型 FM9G4BV 作为对比基线，在统一的遥感多模态评测集 VRSBench 上开展性能对比实验。

实验评估结果表 5 表明，FM9G4BV-RS 在语义质量与空间一致性相关指标上均显著优于基座模型，尤其在 BLEU-4、METEOR 与 ROUGE-L 等反映语义深度与结构连贯性的指标上提升幅度明显，表明遥感领域任务微调与检索增强机制可有效增强模型对细粒度语义的捕捉能力。在空间定位任务中，FM9G4BV-RS 的 IoU@0.5 达到 30.7%，较基座模型提升 22.3 个百分点，说明该框架能够改善视觉 token



(a)



(b)

图 13　评估结果

Fig 13　Evaluation Results

与文本指令之间的空间对齐能力。需要说明的是，本文中的视觉定位指标主要对应 visual grounding 任务，不直接等同于标准目标检测 mAP 评价。综合评分采用归一化换算方式计算：X_1=25×(BLEU-1+BLEU-2+BLEU-4)/(3×100)，X_2=25×Acc@0.5/100，X_3=50×VQA(All)/100，X_4=X_1+X_2+X_3。MME RealWorld RS 仅作为补充评测来源，不参与表 5 中 X_4 总分计算。

#(6)

其中，X_1 表示图像描述/语义生成能力的归一化得分，X_2 表示视觉定位能力得分，X_3 表示视觉问答能力得分，X_4 表示三类任务的综合得分。由表 5 可知，FM9G4BV-RS 的综合得分由基座模型 FM9G4BV 的 15.9 提升至 37.1，增幅较为明显，表明所提出的自适应切片、参数高效微调和遥感知识增强能够有效改善模型在图像描述、视觉定位和视觉问答任务中的整体表现。尽管其绝对得分仍低于 LLaVA-1.5 和 GeoChat，但考虑到本文采用轻量化基座模型，并以较低的训练参数量和显存开销完成遥感领域适配，FM9G4BV-RS 在性能增益、计算成本和部署可行性之间体现出较好的综合优势。

5.3　消融实验与误差分析

为进一步验证 FM9G4BV-RS 中各关键模块的有效性，本文在 VRSBench 测试集上设计组件级消融实验。所有消融设置保持相同的测试样本、图像输入尺度、prompt 模板、最大生成长度和评价脚本，仅改变待分析模块。评价指标包括用于图像描述与语义生成的 BLEU-1、BLEU-2、BLEU-4、METEOR 和 ROUGE-L，用于视觉定位的 Acc@0.5(IoU All)，以及用于遥感视觉问答的 VQA(All)。实验结果如表 6 所示。

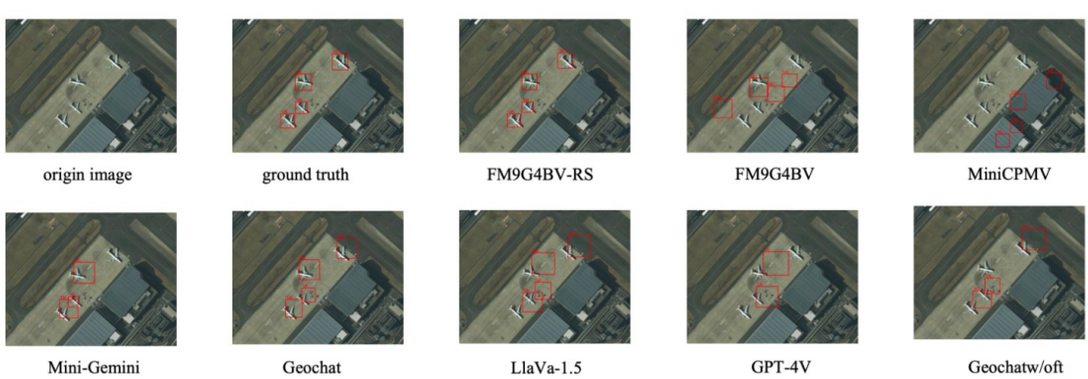


图 14　遥感影像中的多模态视觉接地任务结果对比(一)-飞机

Fig 14　Comparison of Multimodal Visual Grounding Task Results in Remote Sensing Images (Part 1) - Aircraft

由表 6 可见，完整模型在所有语义生成和视觉定位指标上均明显优于未适配的 FM9G4BV 基座模型，说明遥感任务微调和跨模态适配是性能提升的主要来源。其中，去除 LoRA 后各项指标下降最为

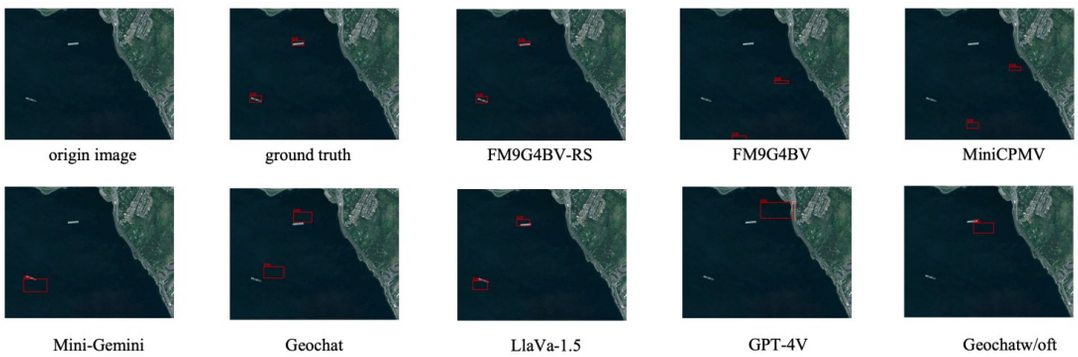


图 15　遥感影像中的多模态视觉接地任务结果对比(二)- 船舶

Fig 15　Comparison of Multimodal Visual Grounding Task Results in Remote Sensing Images (Part 2) - Ship

表 5　模型评估综合评分

Table 5　Model evaluation comprehensive score

Model	X_1	X_2	X_3	X_4
GeoChatw/oft	1.8	3.2	20.4	25.5
GPT-4V	5.7	1.3	32.	39.
MiniGPT-v2	5.7	9.0	19.1	33.
LlaVA-1.5	7.9	10.4	38.	56.
GeoChat	7.6	12.	38.0	58.
Mini-Gemini	7.8	7.5	38.	54.2
MinicpmV	2.1	2.1	11.6	15.
FM9G4BV	2.1	2.1	11.7	15.9
FM9G4BV-RS	5.8	7.7	23.7	37.1

显著，BLEU-4 由 10.1% 降至 3.8%，Acc@0.5 由 30.7% 降至 17.6%，表明仅冻结基座模型难以充分学习遥感图像中的细粒度空间语义。去除 RAG 后，METEOR 和 ROUGE-L 分别下降 3.2 和 2.6，说明外部遥感知识能够改善专业术语表达、灾害场景解释和复杂地物描述。固定 3×3 切片替代自适应切片后，视觉定位指标下降 6.9 个百分点，说明自适应切片对小目标、边界目标和非均匀空间布局具有更好的保持作用。GQA 替换为 MQA 后，各项指标均出现下降，表明适度保留分组键值表示有利于长视觉 token 序列下的跨模态对齐。关闭跨帧时序建模对常规单图指标影响较小，但会削弱双时相变化问答和时间关系表达能力。FP16 推理的精度略高于 NF4 量化版本，但其显存占用和模型存储成本更高，因此 NF4 更适合资源受限平台部署。

5.4　NF4 量化效率分析

由表 7 可知，FP16 配置下模型理论权重显存约

表 6　FM9G4BV-RS 不同模块消融实验结果

Table 6　Ablation Study Results for Different Modules of FM9G4BV-RS

实验设置	BLEU-1	BLEU-2	BLEU-4	Acc@0.5	VQA	METEOR	ROUGE-L
FM9G4BV	16.4	7.6	1.7	8.4	23.3	25.3	17.8
Full	34.8	24.2	10.1	30.7	47.3	35.6	37.2
w/o RAG	32.1	21.6	8.7	29.4	43.8	32.4	34.6
w/o LoRA	22.9	12.4	3.8	17.6	31.5	27.1	23.9
Fixed Slice	31.6	20.5	8.2	23.8	43.1	32.7	33.5
MQA	32.8	21.9	8.8	27.2	44.5	33.1	34.7
w/o Temporal	34.1	23.5	9.6	29.9	45.6	34.8	36.1
FP16	35.0	24.3	10.2	30.9	47.5	35.7	37.4

注：Full 表示完整 FM9G4BV-RS；w/o RAG 表示去除检索增强；w/o LoRA 表示冻结基座模型并关闭参数高效微调；Fixed Slice 表示固定 3×3 切片；MQA 表示将 GQA 替换为多查询注意力；w/o Temporal 表示关闭跨帧时序建模；FP16 表示以全精度推理替代 NF4 量化。

为 7.45 GiB，而 NF4 和 NF4 双重量化分别降低至约 2.10 GiB 和 1.98 GiB，理论压缩率分别达到 71.9% 和 73.4%。从峰值显存看，NF4 配置较 FP16 明显降低，说明低比特权重量化能够有效缓解大模型在显存受限设备上的部署压力。

在性能方面，NF4 配置下 BLEU-4、Acc@0.5 和 VQA（All）相较 FP16 分别下降 0.2、0.3 和 0.3 个百分点，整体变化较小。采用双重量化后，显存进一步下降，但部分指标出现轻微波动。综合来看，普通 NF4 在显存压缩和任务性能之间表现出更好的平衡，因此本文将其作为资源受限平台上的主要量化方案。

表 7　不同量化配置的显存与性能对比

Table 7　Memory and Performance Comparison of Different Quantization Configurations

量化配置	平均存储位宽/(bit·parameter^{-1})	显存/GiB	理论压缩率/%	峰值显存/GiB	BLEU-4/%	Acc@0.5/%	VQA(All)/%
FP16	16.00	7.45	0.0	13.8	10.1	30.7	47.3
NF4	4.50	2.10	71.9	8.1	9.9	30.4	47.0
NF4+双重量化	4.25	1.98	73.4	7.6	9.8	30.2	46.8

结果表明，NF4 能够明显降低模型权重和推理阶段的显存占用，而图像描述、视觉定位和视觉问答指标仅出现小幅变化。采用双重量化后，显存占用进一步下降，但性能相较普通 NF4 略有降低。总体而言，NF4 在模型性能和资源开销之间取得了较好的平衡，适合显存受限条件下的遥感多模态模型部署。

6　结　论

本文围绕高分辨率遥感影像在通用多模态大模型中存在的长序列编码开销大、细粒度空间信息易损失和领域知识不足等问题，提出了一种基于检索增强与参数高效微调的遥感影像多任务语义理解框架 FM9G4BV-RS。该框架以 FM9G4BV 为基座模型，通过自适应高分辨率切片实现大幅面图像的结构保持性编码，结合 GQA 提升长视觉 token 序列下的跨模态对齐效率，并利用 LoRA 微调与 NF4 量化在控制训练和部署成本的同时完成遥感任务适配。同时，本文引入 RAG 机制，将遥感文献、数据集说明和灾害报告等外部知识作为可追溯上下文融入模型推理过程，从而增强专业遥感语义生成和复杂场景解释能力。

在 VRSBench 视觉语言评测基准上的实验结果表明，FM9G4BV-RS 在图像描述、视觉定位和遥感视觉问答任务上均较基座模型取得明显提升。模型的 BLEU-1、BLEU-2 和 BLEU-4 分别达到 34.8%、24.2% 和 10.1%，METEOR 和 ROUGE-L 分别达到 35.6 和 37.2，Acc@0.5（IoU All）达到 30.7%，VQA（All）达到 47.3%。与未适配的 FM9G4BV 相比，本文方法在文本生成质量、空间定位能力和问答准确性方面均表现出更强的遥感场景适应性。消融实验进一步表明，LoRA 微调是模型领域适配性能提升的关键因素，RAG 主要改善专业语义表达和知识一致性，自适应切片对小目标定位和边界区域保持具有重要作用，GQA 和 NF4 量化则在长序列建模效率和轻量化部署方面具有实际价值。

需要指出的是，本文仍存在一定局限性：一方面，部分变化检测、地物分类和灾害评估任务仍以框架支持和案例分析为主，后续仍需在 FAIR1M、NWPU VHR-10、LEVIR-CC、BRIGHT 和 xBD 等专项数据集上补充更系统的标准化定量实验；另一方面，模型在密集小目标、极小尺度地物、跨季节伪变化和复杂背景干扰场景中仍可能出现定位偏移或语义混淆。未来工作将进一步扩展多源遥感数据训练规模，引入更细粒度的时空对齐机制和不确定性评估方法，探索端云协同部署下的实时遥感智能解译应用，并可结合大模型-智能体协同技术构建动态可扩展的交互式解译系统（Qi 等，2026）。

参考文献（References）

Wang G H, Zhang T, Liu Y, et al. 2025. Development of large models for intelligent remote sensing interpretation and prospects for their application in cultivated land change detection. Land and

Resources Herald, 22(04): 10-23. (王光辉,张涛,刘宇,等.遥感智能解译大模型发展及其耕地变化检测应用展望[J].国土资源导刊,2025,22(04):10-23).

Zheng X T, Xiao X L, Chen X M, et al. 2024. Research Progress on Cross-Domain Remote Sensing Scene Interpretation. Journal of Chinese Journal of Image and Graphics, 29(6):1730-1746. (郑向涛,肖欣林,陈秀妹,等.跨域遥感场景解译研究进展[J].中国图象图形学报,2024,29(6):1730-1746).

Chen H, Song J, Dietrich O, et al. 2025. BRIGHT: A globally distributed multimodal building damage assessment dataset with very-high-resolution for all-weather disaster response. Earth System Science Data, 17(11): 6217-6251. [DOI: 10.5194/essd-17-6217-2025]

Yan K, Xu J M and Wang Q. 2025. Lightweight large-model method for remote sensing detection of surface anomalies. Acta Geodaetica et Cartographica Sinica, 54(09): 1664-1676. (闫凯,徐健明,王桥.地表异常遥感探测轻法[J].测绘学报,2025,54(09):1664-1676).

Wei Y Y, Mao T Y, Li B A, et al. 2025. Research Progress on Visual Models and Multimodal Large Models Promoting Image Restoration and Enhancement. Journal of Image and Graphics in China, 30(5): 1197-1219. (韦炎炎,毛天一,李柏昂,等.视觉模型及多模态大模型推进图像复原增强研究进展[J].中国图象图形学报,2025,30(5):1197-1219).

Dang Y, Zhu M, Wang D, et al. 2025. A Benchmark for Ultra-High-Resolution Remote Sensing MLLMs. arXiv preprint: arXiv: 2512.17319.

Dong S, Wang L, Du B, et al. 2024. ChangeCLIP: Remote sensing change detection with multimodal vision-language representation learning. ISPRS Journal of Photogrammetry and Remote Sensing, 208: 53-69. [DOI: 10.1016/j.isprsjprs.2024.01.015]

Wang W J, Yang W H, Fang Y M, et al. 2024. A Review of Visual Perception and Understanding in Harsh Scenes. Journal of Image and Graphics of China, 29(6):1667-1684. (汪文靖,杨文瀚,方玉明,等.恶劣场景下视觉感知与理解综述[J].中国图象图形学报,2024,29(6):1667-1684).

Feng Q L, Chen B A, Li G Q, et al. 2022. A review of remote sensing image sample datasets. Journal of Remote Sensing, 26(4): 589-605. (冯权泷, 陈泊安, 李国庆, 等. 2022. 遥感影像样本数据集研究综述. 遥感学报, 26(4): 589-605). [DOI: 10.11834/jrs.20220045]

Ge J, Zhang X, Zheng Y, et al. 2025. RSTeller: Scaling up visual language modeling in remote sensing with rich linguistic semantics from openly available data and large language models. ISPRS Journal of Photogrammetry and Remote Sensing, 226: 146-163. [DOI: 10.1016/j.isprsjprs.2025.04.012]

Li Q, Ma S, Luo J, et al. 2026. Co-Training Vision-Language Models for Remote Sensing Multi-Task Learning. Remote Sensing, 18(2): 222. [DOI: 10.3390/rs18020222]

Li X, Ding J, Elhoseiny M. 2024. Vrsbench: A versatile vision-language benchmark dataset for remote sensing image understanding. Advances in Neural Information Processing Systems, 37: 3229-3242.

Liu C, Zhao R, Chen H, et al. 2022. Remote sensing image change captioning with dual-branch transformers: A new method and a large scale dataset. IEEE Transactions on Geoscience and Remote Sensing, 60: 1-20. [DOI: 10.1109/TGRS.2022.3148765]

Lu X, Wang B, Zheng X, et al. 2017. Exploring models and data for remote sensing image caption generation. IEEE Transactions on Geoscience and Remote Sensing, 56(4): 2183-2195. [DOI: 10.1109/TGRS.2017.2776321]

Ma W L, Sun T, Zhao R X, et al. 2026. Research on graphical abstract generation of scientific and technological literature based on LLM and GraphRAG. Data Analysis and Knowledge Discovery, 1-21. (马玮璐, 孙坦, 赵瑞雪, 等. 2026. 基于LLM和GraphRAG的科技文献图谱式摘要生成研究. 数据分析与知识发现, 1-21). [DOI: 10.11925/infotech.2096-3467.2025.0345]

Qin Y F, Fu X, Zhang Z X, et al. 2025. Key technologies for the application of coal mine professional large model based on LoRA fine-tuning and RAG fusion. Industry and Mine Automation, 51(8): 34-42, 50. (秦一凡, 付翔, 张智星, 等. 2025. 基于LoRA微调与RAG融合的煤矿专业大模型应用关键技术. 工矿自动化, 51(8): 34-42, 50). [DOI: 10.13272/j.issn.1671-251x.2025030031]

Soni S, Dudhane A, Debary H, et al. 2025. Earthdial: Turning multi-sensory earth observations to interactive dialogues. Proceedings of the Computer Vision and Pattern Recognition Conference, 14303-14313.

Tao L, Zhang H, Jing H, et al. 2025. Advancements in vision - language models for remote sensing: Datasets, capabilities, and enhancement techniques. Remote Sensing, 17(1): 162. [DOI: 10.3390/rs17010162]

Zhang S, Shan L, Qiu R. 2025. Multimodal Interpretation of Remote Sensing Images: Dynamic Resolution Input Strategy and Multi-scale Vision-Language Alignment Mechanism. arXiv preprint: arXiv:2512.23243.

Zhang W, Cai M, Zhang T, et al. 2024. EarthGPT: A universal multi-modal large language model for multisensor image comprehension in remote sensing domain. IEEE Transactions on Geoscience and Remote Sensing, 62: 1-20. [DOI: 10.1109/TGRS.2024.3387654]

Zhou G, Lihuang Q, Gamba P. 2025. Advances on multimodal remote sensing foundation models for Earth observation downstream tasks: A survey. Remote Sensing, 17(21): 3532. [DOI: 10.3390/rs17213532]

Qi X Y, Shi H R, Wu Y F, et al. 2026. Research on a dynamically scalable intelligent remote sensing target interpretation system driven by collaboration between large models and agents. Journal of Aerospace Engineering University, 3(02): 36-46. (齐析屿,师汉儒,吴一凡,等.大模型-智能体协同驱动的动态可扩展遥感目标智能

解译系统研究[J].航天工程大学学报,2026,3(02):36-46).

作者简介

张婧楠,女,本科生,主要研究方向为计算机视觉。E-mail:zjnjingnan@163.com

张逢骏,通信作者,男,讲师,主要研究方向为三维计算机视觉。E-mail:jszhfi@nuaa.edu.cn