# 面向遥感智能体的多模态图文指令大规模数据集

王佩瑾[①②③④] 胡会扬*[①②③④] 冯瑛超[①④] 刁文辉[①②③④] 孙 显[①②③④]

[①](中国科学院空天信息创新研究院 北京 100190)

[②](中国科学院大学 北京 100190)

[③](中国科学院大学电子电气与通信工程学院 北京 100049)

[④](目标认知与应用技术国家级重点实验室 北京 100190)

**摘 要**：随着遥感应用不断从静态图像分析迈向智能化认知决策任务，构建覆盖多任务、多模态的信息融合数据体系已成为推动遥感基础模型发展的关键前提。该文围绕遥感智能体中的感知、认知需求，提出并构建了一个面向多任务图文指令的遥感多模态数据集，系统组织图像、文本指令、空间坐标与行为轨迹等多模态信息，统一支撑多阶段任务链路的训练与评估。该数据集涵盖9类核心任务，包括关系推理、指令分解、任务调度、定位描述与多模态感知等，共计21个子数据集，覆盖光学、SAR与红外3类遥感模态，总体数据规模超过2 000 000样本。在数据构建过程中，该文针对遥感图像的特性设计了标准化的指令格式，提出统一的输入输出范式，确保不同任务间的互通性与可迁移性。同时，设计自动化数据生成与转换流程，以提升多模态样本生成效率与一致性。此外，该文还介绍了在遥感基础模型上的基线性能评估结果，验证了该数据集在多任务泛化学习中的实用价值。该数据集可广泛服务于遥感领域多模态基础模型的构建与评估，尤其适用于统一感知-认知-决策闭环流程的智能体模型开发，具有良好的研究推广价值与工程应用前景。

**关键词**：遥感基础模型；多模态指令数据集；感知-认知-决策

**中图分类号**：TN911.7 **文献标识码**：A **文章编号**：1009-5896(2026)04-1608-15

**DOI**: 10.11999/JEIT250818 **CSTR**: 32379.14.JEIT250818

## 1 引言

遥感作为对地观测的重要手段，在城市监测、环境评估、农业生产、灾害应急等关键领域发挥着不可替代的作用。伴随对遥感信息理解深度和决策效率的不断提升需求，遥感技术正逐步从传统的“感知识别”向“认知推理”转型。从早期依赖人工特征和浅层分类器的遥感识别方法，到近年来广泛应用的深度学习模型，遥感图像的解析能力已取得显著提升[1–3]。然而，当前主流研究多集中在孤立的感知任务上[4–6]，如地物分类、目标检测、变化检测等，模型能力往往局限于“看到什么”，缺乏面向“如何理解、如何执行”的高级推理与交互能力。

多模态基础模型(Multimodal Foundation Models, MFMs)为解决这一问题提供了新的范式。以MiniGPT-v2[7], Kosmos-2[8], Shikra[9] 等为代表的多模态基础模型已在自然图像领域展现出良好的通用视觉理解与语言表达能力，引发图文对预训练、跨模态问答、指令生成等方向的快速发展。然而，直接将这些方法迁移至遥感领域面临显著困难：一方面，遥感图像具有幅宽大、场景复杂、目标尺寸变化大等特征，直接迁移泛化性差；另一方面，遥感应用所需任务链更长、语义更复杂，往往涉及目标感知、空间推理、任务规划与反馈交互等完整链路。构建具备任务泛化能力的遥感多模态基础模型，亟需支持感知、认知、推理和执行等多任务、多阶段的高质量训练数据。

近年来，研究者开始尝试构建遥感多模态数据集以支持上述目标。LEVIR-CC[10], RSVQA[11], NWPU-Captions[12], RSICD[13]和UCM-Captions[14]等数据集引入图文描述任务，推动了遥感图文生成与跨模态检索研究的发展；HRVQA[15]数据集则构建了遥感视觉问答任务，探索图文联合推理能力。这些数据集在一定程度上拓宽了遥感多模态研究的边界，为领域早期的发展奠定了基础。

随着任务类型不断演进，面向视觉语言模型的多任务数据集也逐渐涌现，为统一、综合的模型评估提供了新的可能。例如，RSIEval[16]将DOTA-v1.5的部分图像裁剪为512×512的图像块，并由专家进行精细标注，构建了包含100个高质量图像-描述对与936个图像-问答3元组的评测数据集，为图像描

收稿日期：2025-08-29；改回日期：2026-01-11；网络出版：2026-01-27
*通信作者： 胡会扬 huhuiyang22@mails.ucas.ac.cn
基金项目：国家重点研发计划(2024YFF1401001)，中国科学院空天信息创新研究院科学与颠覆性技术项目(2025-AIRCAS-SDTP-04)
Foundation Items: The National Key R&D Program of China (2024YFF1401001), The Science and Disruptive Technology Program, AIRCAS (2025-AIRCAS-SDTP-04)

述与视觉问答任务提供了多角度的评估基准。SkySenseGPT[17]提出了规模达1 800.8 k的大型多任务数据集FIT-RS，覆盖描述、问答、检测和推理等多种全局与区域级理解任务，旨在促进模型对复杂遥感场景中丰富语义关系的深层次建模。VRSBench[18]则构建了一个包含29 614条详细图像描述、52 472条目标引用以及123 221条问答对的视觉语言基准，支持图像描述、视觉定位与视觉问答等核心任务。

然而，现有数据集仍存在较为明显的局限性。首先，任务形式仍主要集中于静态感知任务，缺乏多轮交互推理、多模态协同等更复杂的认知类任务设置；其次，多模态覆盖有限，绝大多数数据集仅支持光学遥感影像，而缺乏SAR、红外等多传感器信息的整合；此外，一些数据集尚未完全开源，或规模偏小，难以满足大规模视觉语言模型的预训练需求。

为推动遥感领域的“感知-认知-决策”一体化建模进程，亟需构建一个任务覆盖广、模态结构丰富、标注方式多样、能够支持模型训练与评估的大规模开放数据集。为此，基于遥感智能解译领域的若干公开数据集，本文构建了一个统一面向遥感感知与认知一体化任务的数据集体系。

本数据集在内容上具备如下关键特征：一是任务链条完整，涵盖图像描述、目标检测、问答推理、空间定位和指令分解等全流程任务场景，支持遥感大模型从静态感知走向动态决策；二是模态和平台类型多样，除光学卫星图像和文本外，还引入SAR、红外等多传感器、无人机等多平台数据，提升模型在空间认知和场景理解上的能力；三是样本组织标准统一，采用图文指令格式构建多任务样本，适用于多模态预训练、指令微调与统一评估训练；四是数据量大且开放共享，总计200多万多任务指令样本，能够支撑不同规模与架构的遥感多模态模型训练。

综上所述，本文提出的遥感图文指令数据集不仅是对现有数据集的一次全面拓展，更是面向遥感智能体系统演进需求的数据基础设施建设。后续章节将详细介绍该数据集的构建方法、任务覆盖、可视化展示、评价方法及其在遥感大模型中的应用验证。

## 2　数据集属性

本数据集涵盖9个任务场景和3种感知模态，总计21个子数据集。每个子数据集对应一个独立文件夹，文件夹内的数据以JSON格式存储，图像文件为JPG或PNG格式。每个JSON文件包含4类标签信息：“image”表示图像路径，“text_input”为指令输入，“text_output”为模型输出答案，“image_id”为每条数据的唯一标识。数据集整体内容及分布情况如表1所示。

## 3　数据集描述

### 3.1　任务1：关系推理

关系推理任务旨在超越单一目标检测或分类的范畴，进一步建模目标之间的语义关联。该任务要求模型在识别图像中主体与客体类别的基础上，推断二者的空间或功能关系，从而实现更高层次的场景语义理解。与仅依赖目标类别信息的传统视觉关系检测不同，本文将关系推理视为一种由自然语言指令驱动的推理任务：模型不仅回答“有什么”，还需根据指令执行“如何推理目标间的语义关系”。

#### 3.1.1　数据来源

本任务所使用的数据来源于Yan等人[24]提出的遥感关系理解数据集ReCon1M，并在此基础上构建关系推理任务数据。原始数据集包含21 392 张遥感图像，空间分辨率范围为0.3～0.8 m，标注了60个类别中的859 751个目标边界框，并基于这些标注生成了涵盖64类关系的1 149 342组关系3元组。ReCon1M数据仅作为基础视觉标注来源，而

表 1　数据集整体统计

| 任务类型 | 数据模态 | 数据集名称 | 数量规模 |
|---|---|---|---|
| 任务调度 | 光学 | Citynav[19] | 32637 |
| 指令分解 | 光学 | ReCon1M-DEC | 27821 |
| 关系推理 | 光学 | ReCon1M-REL | 125000 |
| 关系检测 | 光学 | ReCon1M-DET | 120097 |
| 定位描述 | 光学 | DIOR-GC | 22921 |
| | 光学 | DOTA-GC | 48866 |
| 多模态感知(目标检测) | 光学 | DIOR[20] | 23463 |
| | SAR | SARDet-100K[21] | 116598 |
| | 红外 | IR-DET | 56353 |
| 多模态感知(图像描述) | 光学 | DIOR-CAP | 92875 |
| | 光学 | DOTA-CAP | 307150 |
| | SAR | SAR-CAP | 582990 |
| | 红外 | IR-CAP | 281765 |
| 多模态感知(图像分类) | 光学 | AID[22] | 10000 |
| | 光学 | NWPU-RESISC45[23] | 31500 |
| | SAR | SAR-CLA | 116597 |
| | 红外 | IR-CLA | 56353 |
| 多模态感知(目标计数) | 光学 | DIOR-COUNT | 35204 |
| | 光学 | DOTA-COUNT | 78432 |
| | SAR | IR-COUNT | 107565 |
| | 红外 | SAR-COUNT | 117803 |
| 总数 | - | - | 2391990 |

本文通过重新定义任务形式实现了从“关系标注数据”到“关系推理任务数据”的转化。

#### 3.1.2 任务设置

与传统视觉关系检测方法不同，本文提出的关系推理任务通过明确的指令模板、结构化输出以及任务显式标识来实现推理能力，是一种较新的任务形式。在本文中，模型的输入由一条自然语言指令和一幅遥感图像组成，指令明确指定了两个待推理目标的空间位置及其关系推断要求。如图1所示，指令要求模型推断位于坐标[[473,116,546,132]]的主体(warship)与位于坐标[[0,0,997,166]]的客体之间的关系，并输出两者的类别标签。其中，<|reasoning|> 用于显式指示该指令属于关系推理任务，使多任务模型能够区分不同任务类型并调用相应的推理能力；<|det|>…<|/det|>用于标注目标在图像中的边界框位置，坐标已归一化至[0, 999]范围，以适配不同分辨率图像的统一表示；<|ref|>…<|/ref|>用于显式指明已知类别的目标；<|rel|>…<|/rel|>标签用于标注关系类别。这些标签均为本文提出的统一指令框架中的组成部分，不属于 ReCon1M 数据集原生内容。

模型的输出包括3个要素：(1)主体类别，(2)客体类别，(3)二者的语义关系，如图1所示。该格式既保证了输出的结构化表示，便于自动化解析，构成统一的结构化预测格式。

#### 3.1.3 数据集制作流程

本文基于ReCon1M数据集的原始目标与关系标注，重新定义了任务样本的组织方式、指令输入结构与模型可解析的输出格式，并据此构建了全新的关系推理数据集 ReCon1M-REL。在目标几何表示方面，针对原始标注中的多边形目标框，本文设计了统一的坐标规范化机制，将坐标值按图像宽高分别归一化至[0,999]的整数区间。其次，在样本构建方面，根据本文提出的关系推理任务定义重新组织样本结构：为每个主体-客体对生成一条独立的指令响应样本，显式绑定指令内容与期望输出。

最后，在划分策略上，本文限制数据总量为125 000条样本，并按照9:1的比例随机划分为训练集与测试集，最终得到训练集与测试集文件分别为112 500和12 500条。

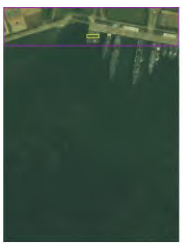



图 1 关系推理任务问答对示例

#### 3.1.4 语言描述质量评估分析

在该任务中，由于问答模板设置固定，类别关系的样本分布直接影响模型对各类关系的学习效果与泛化能力。表2展示了关系推理数据集中各类别关系的样本分布情况。整体来看，该数据呈现出典型的长尾分布特征：少数关系类别出现频率极高，而大量类别的样本数量相对稀少。首先，从高频关系类别来看，如close-to, park-next-to, parked-at以及 provide-access-to等类别占据了数据集的主要比例。这些关系通常涉及空间邻近、停泊、通行等基本场景语义，是遥感图像中普遍出现的关系类型。相比之下，低频关系类别数量较多，并且许多类别在训练集中的样本仅有个位数，例如dig, emit, pass-under, ventilate等。这对模型的学习能力提出挑战，具有对于小样本关系推理的适配价值。在类别覆盖方面，训练集覆盖了59种关系种类，测试集包含其中的50种。同时，训练集包含主体类别50个、客体类别51个，测试集包含主体类别47个、客体类别48个。

### 3.2 任务2：关系检测

关系检测任务旨在识别图像中目标实体之间的语义关联及空间联系。该任务不仅要求模型准确定位目标的位置，还需推断它们之间的关系类别，从而实现更高层次的场景理解。

#### 3.2.1 数据来源

本任务的数据同样构建于Yan等人[24]提出的ReCon1M遥感关系理解数据集，但本文对其进行了面向检测任务的重新组织与样本重构。通过引入统一坐标体系、指令化输入模板以及结构化输出格式，原始标注被转化为适用于关系检测的指令响应样本，使模型能够在本文统一的多任务框架下执行关系检测。

#### 3.2.2 任务设置

模型的输入由一条自然语言指令和一幅遥感图像组成，指令明确指定需要分析的两个目标实体及其位置关系。指令采用如图2所示的格式，其中<|detection|>用于显式指示该指令属于关系检测任务，使多任务模型能够区分不同任务类型；<|ref|>，<|det|>和<|rel|>等标签作为本文采用的跨任务统一指令体系的一部分，含义与上文保持一致。模型的输出为结构化的关系预测结果，表示目标对的空间关系。

#### 3.2.3 数据集制作流程

本研究所使用的关系检测数据集基于ReCon1M，依据本文提出的指令式任务框架进行了面向检测任

表 2　关系推理数据集中各类别关系的样本分布统计表

| 类别名 | 训练集 | 测试集 | 类别名 | 训练集 | 测试集 | 类别名 | 训练集 | 测试集 | 类别名 | 训练集 | 测试集 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| above | 69 | 5 | dig | 1 | 0 | link-to | 166 | 21 | pull | 5 | 1 |
| adjacent-to | 2799 | 289 | dock-alone-at | 4 | 1 | load | 9 | 0 | sail-by | 20 | 4 |
| adjoint-with | 3 | 1 | dock-at | 102 | 13 | manage | 35 | 4 | sail-on | 922 | 123 |
| around | 3 | 0 | drive-at-the-different-lane | 148 | 20 | moor-at | 3414 | 383 | separate | 31 | 1 |
| belong-to | 913 | 112 | drive-at-the-same-lane | 144 | 17 | move-away-from | 8 | 0 | serve | 1737 | 174 |
| block | 2 | 1 | drive-on | 4931 | 538 | park-alone-at | 14 | 3 | slow-down | 241 | 34 |
| border | 72 | 9 | emit | 1 | 0 | park-next-to | 30544 | 3389 | supplement | 334 | 50 |
| close-to | 24992 | 2657 | enter | 6 | 0 | parked-at | 15432 | 1735 | supply | 179 | 26 |
| command | 3 | 0 | equipped-with | 66 | 8 | pass-under | 3 | 0 | support | 79 | 7 |
| connect | 186 | 15 | exit-from | 5 | 1 | pile-up-at | 660 | 83 | support-the-construction-of | 82 | 9 |
| contain | 106 | 14 | hoist | 93 | 10 | placed-on | 29 | 2 | taxi-on | 226 | 22 |
| converge | 14 | 1 | inside | 7558 | 854 | power | 139 | 9 | tow | 10 | 3 |
| cooperate-with | 430 | 47 | is-parallel-to | 3516 | 402 | prepared-for | 81 | 14 | transport | 48 | 6 |
| cross | 1224 | 130 | is-symmetric-with | 4 | 1 | provide-access-to | 10584 | 1247 | ventilate | 44 | 2 |
| cultivate | 10 | 0 | lie-under | 13 | 1 | provide-shuttle-service-to | 6 | 1 | | | |


图 2 关系检测任务问答对示例

务的重构与语义规范化。首先，从ReCon1M中筛选具有明确主体-客体实体及其关系语义的样本，确保图像中包含可用于关系检测的有效目标对。随后，对原始标注进行重新组织，包括统一坐标体系、重设主体/客体引用方式以及构建可被模型解析的结构化指令模板。最终构建的数据集涵盖多类典型关系，命名为ReCon1M-DET。数据共有120 097条样本，并按9:1的比例随机划分为训练集与测试集，得到训练集108 057条，测试集12 040条。

#### 3.2.4　语言描述质量评估分析

本文对数据集中的关系类别在训练集与测试集中的分布进行了统计分析。如表3结果显示，训练集共包含55个关系类别，测试集包含42个关系类别，且测试集中所有类别均已在训练集中出现。数据集中存在明显的长尾分布现象。少数关系类别如park-next-to, close-to和parked-at，拥有极其丰富的训练样本，构成了数据分布的主体。

### 3.3　任务3：指令分解

为提升模型对复杂遥感指令的理解与执行能力，本文创新性地提出了指令分解任务。该任务要求模型将一条高层次的、可能涉及特定区域或语义目标的自然语言指令，系统地拆解为一个结构化的、可解释的多步推理序列。

#### 3.3.1　数据来源

在ReCon1M遥感关系理解数据集的基础上，进行了任务重构与定义创新。与ReCon1M侧重于单一的关系检测不同，本文的任务定义具有显著的步骤性：本文将指令理解构建为一个包含空间定位、目标检测、关系推理与上下文总结的序列化决策过程。这不仅扩展了原始数据的应用范畴，更引入了一种全新的、面向任务可解释性的复杂指令理解范式。

#### 3.3.2　任务设置

在指令分解任务中，本文设计了一套统一的指令模板与严格结构化的输出格式。模型的输入由一条自然语言指令和一幅遥感图像组成，指令明确指定了需要分析的目标区域。如图3所示，该任务要求模型分析位于坐标[[627,82,739,326]]的图像区

域，并依次完成区域定位、目标检测、关系分析及上下文总结等子任务。其中，<|decomposition|>用于显式指示该指令属于指令分解任务，使多任务模型能够区分不同任务类型并调用对应的推理能力；其他标签的作用与上文一致。

模型的输出为结构化的多步骤推理结果，涵盖4个要素：(1)目标区域的空间位置描述；(2)区域内所有检测目标的类别与位置；(3)目标间的语义关系及对应的参与目标；(4)整体上下文总结，包括检测目标数量、类别数及关系数。该格式保证输出的可解释性和结构化表示，便于自动化解析与评估。

### 3.3.3 数据集制作流程

本文基于ReCon1M数据集的原始目标与关系标注，构建了自动化的数据转换与格式化流程。关系的提取方式与前文保持一致，同时将每幅图像划分为九个区域，用于判定目标所属的空间范围，并将结果组织为结构化输出，新数据集命名为ReCon1M-DEC。在数据划分阶段，本文将数据总量限定为27 821条样本，并按8:2的比例随机划分为训练集与测试集，最终得到训练集22 256条，测试集5 565条。

### 3.3.4 语言描述质量评估分析

同样地，由于模板设置的固定性，为了科学评估任务质量，需关注类别数量和关系数量的分布情况，因为这些分布直接反映了数据的覆盖广度和多样性，进而影响模型的学习效果和泛化能力。如表4所示，在数据划分结果中，训练集共覆盖59种关系类别，测试集则包含其中的56种，且测试集中所有关系类别均已在训练集中出现。

## 3.4 任务4：任务调度

为探索大模型在复杂城市环境中进行自主空间决策的能力，本文设计了任务调度任务。该任务模

表 3 关系检测数据集中各类别关系的样本分布统计表

| 类别名 | 训练集 | 测试集 | 类别名 | 训练集 | 测试集 | 类别名 | 训练集 | 测试集 | 类别名 | 训练集 | 测试集 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| above | 73 | 7 | dock-alone-at | 5 | 0 | manage | 29 | 5 | sail-by | 22 | 3 |
| adjacent-to | 2575 | 297 | dock-at | 80 | 7 | moor-at | 3224 | 344 | sail-on | 636 | 78 |
| adjoint-with | 4 | 0 | drive-at-the-different-lane | 112 | 15 | move-away-from | 5 | 0 | separate | 39 | 4 |
| around | 4 | 2 | drive-at-the-same-lane | 154 | 18 | park-alone-at | 10 | 0 | serve | 1655 | 186 |
| belong-to | 768 | 66 | drive-on | 4819 | 545 | park-next-to | 29857 | 3249 | slow-down | 245 | 34 |
| block | 1 | 0 | enter | 8 | 0 | parked-at | 15308 | 1729 | supplement | 321 | 37 |
| border | 52 | 7 | equipped-with | 59 | 9 | pass-under | 3 | 0 | supply | 207 | 16 |
| close-to | 23942 | 2713 | exit-from | 5 | 0 | pile-up-at | 590 | 76 | support | 61 | 6 |
| command | 2 | 0 | hoist | 115 | 7 | placed-on | 25 | 0 | support-the-construction-of | 11 | 1 |
| connect | 169 | 16 | inside | 7556 | 860 | power | 99 | 9 | taxi-on | 155 | 13 |
| contain | 111 | 7 | is-parallel-to | 3088 | 345 | prepared-for | 64 | 11 | tow | 7 | 0 |
| converge | 9 | 2 | is-symmetric-with | 2 | 0 | provide-access-to | 9989 | 1136 | transport | 28 | 2 |
| cooperate-with | 384 | 39 | lie-under | 14 | 4 | provide-shuttle-service-to | 5 | 0 | ventilate | 30 | 2 |
| cross | 1151 | 118 | link-to | 168 | 14 | pull | 2 | 1 | | | |

指令：<|decomposition|>Analyze region <|det|>[[627,82,739,326]]<|/det|> of the image.
答案：Step1: Locate the target area: The target area locates at upper right of the image; Step2: Perform object detection: There are 2 objects in the target area, including: 2 <|ref|>airplanes<|/ref|><|det|>[[668,103,731,183],[637,191,727,306]]<|/det|>; Step3: Perform relation analysis: There are 2 relations found: the <|ref|>airplane<|/ref|><|det|>[[668,103,731,183]]<|/det|> is <|rel|>park-next-to<|/rel|> the <|ref|>airplane<|/ref|><|det|>[[637,191,727,306]]<|/det|>, the <|ref|>airplane<|/ref|><|det|>[[637,191,727,306]]<|/det|> is <|rel|>park-next-to<|/rel|> the <|ref|>airplane<|/ref|><|det|>[[668,103,731,183]]<|/det|>; Step4: Perform context summary: Detected 2 objects (1 class) with 2 interactions.

图 3 指令分解任务问答对示例

拟了自主空中飞行器基于遥感影像与自然语言指令进行三维轨迹规划的完整流程，要求模型综合理解地理空间、目标描述与初始状态，生成可执行的三维飞行路径。

### 3.4.1 数据来源

本任务的设定不同于传统的点对点导航，将其定义为一个基于多模态感知的序列决策问题：模型在接收包含目标位置周边空间环境的语言指令及智能体初始状态后，需依次完成地标解析、目标定位与三维轨迹生成。本任务基于CityNav[19]数据集，该数据集专为空中视觉语言任务构建。数据集中包含32 637条人工示范轨迹，每条轨迹均配有对应的自然语言描述，覆盖英国剑桥和伯明翰两座城市共约4.65 $km^2$的真实区域，为模型在复杂城市环境中的轨迹规划提供了丰富的训练和评估资源。本文的主要改进在于对该数据进行了面向端到端决策的任务重构，并设计了相应的指令与输出规范。

### 3.4.2 任务设置

本文为此任务专门设计了一套结构化的指令模板与输出格式，以明确指导模型行为并规范其决策过程。模型的输入整合了遥感图像、自然语言指令及智能体初始状态(三维坐标与姿态角)。如图4所示，指令统一格式为<|navigation|>，模型需根据描述制定飞行计划，使其能够覆盖沿途建筑并抵达目标位置。指令中通过<|ref|>…<|/ref|>明确标识参考地标，<|pos|>…<|/pos|>指定三维坐标(已归一化至[0, 999]范围)，模型需据此确定目标位置及周边环境信息。模型的输出被严格定义为必须依次执行的4个步骤，形成可解释的决策链条：步骤1提取目标及地标信息；步骤2获取各地标三维坐标；

表 4　指令分解数据集中各类别关系的样本分布统计表

| 类别名 | 训练集 | 测试集 | 类别名 | 训练集 | 测试集 | 类别名 | 训练集 | 测试集 | 类别名 | 训练集 | 测试集 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| above | 195 | 48 | dig | 32 | 7 | link-to | 78 | 24 | pull | 1 | 0 |
| adjacent-to | 7960 | 2014 | dock-alone-at | 22 | 5 | load | 70 | 21 | sail-by | 95 | 11 |
| adjoint-with | 20 | 4 | dock-at | 172 | 34 | manage | 108 | 20 | sail-on | 4750 | 890 |
| around | 10 | 5 | drive-at-the-different-lane | 210 | 48 | moor-at | 10740 | 2751 | separate | 14 | 4 |
| belong-to | 3957 | 1108 | drive-at-the-same-lane | 42 | 18 | move-away-from | 35 | 3 | serve | 2623 | 681 |
| block | 11 | 0 | drive-on | 9148 | 2119 | park-alone-at | 41 | 8 | slow-down | 417 | 103 |
| border | 319 | 97 | emit | 12 | 4 | park-next-to | 71926 | 17232 | supplement | 475 | 124 |
| close-to | 40840 | 9980 | enter | 30 | 4 | parked-at | 35623 | 8641 | supply | 403 | 115 |
| command | 34 | 9 | equipped-with | 89 | 21 | pass-under | 28 | 8 | support | 94 | 38 |
| connect | 618 | 170 | exit-from | 6 | 0 | pile-up-at | 2070 | 514 | support-the-construction-of | 61 | 9 |
| contain | 441 | 25 | hoist | 266 | 66 | placed-on | 139 | 20 | taxi-on | 1175 | 320 |
| converge | 17 | 5 | inside | 12566 | 3607 | power | 212 | 47 | tow | 48 | 10 |
| cooperate-with | 998 | 230 | is-parallel-to | 9428 | 2578 | prepared-for | 139 | 31 | transport | 195 | 42 |
| cross | 1534 | 364 | is-symmetric-with | 4 | 4 | provide-access-to | 12362 | 3009 | ventilate | 83 | 25 |
| cultivate | 97 | 28 | lie-under | 42 | 9 | provide-shuttle-service-to | 44 | 34 | | | |


指令：<|navigation|>You need to formulate a flight plan for a quadcopter based on this map, enabling it to fly over all the buildings and reach the destination. The target location is described as follows: This is a black car parked in the parking space near Perry Barr. There is one empty spot on its left. Over this empty spot, there is another black car. The three-dimensional coordinates of the landmark are as follows: <|ref|>Perry Barr<|/ref|><|pos|>[512, 720, 35]<|/pos|>. Your starting 3D coordinates and orientation angles are [444, 234, 410, 572, 143, 156]. You need to provide a series of three-dimensional waypoints and attitude angles for the quadcopter to reach the target location.
答案：Step 1: Extract basic information as follows: Target: black car. Landmarks: Perry Barr. Surroundings: parking space.
Step 2: Get landmarks position: <|ref|>Perry Barr<|/ref|><|pos|>[512, 720, 35]<|/pos|>.
Step 3: Get target position: <|ref|>black car<|/ref|><|pos|>[509, 757, 39]<|/pos|>.
Step 4: Trajectory: <|pose|>[[444, 234, 410, 572, 143, 156], ..., [499, 757, 49, 887, 477, 185]]<|/pose|>.

图 4 任务调度任务问答对示例

步骤3确定目标位置；步骤4生成飞行轨迹，即一系列三维航路点和对应姿态角，使用<|pose|>⋯<|/pose|>标签记录。

#### 3.4.3 数据集制作流程

本任务的数据集源自CityNav[19]，其原始数据主要由离散的飞行轨迹点集和与之对应的标志物描述文本构成。为将其转化为适用于学习动态规划决策的样本，本文设计了一套系统的结构化重构流程，将静态的轨迹数据序列化为一个模拟智能体实时决策的规划过程。具体的构建流程如下：首先，对每条轨迹进行预处理，将初始姿态、三维坐标以及沿途航路点统一记录，并将地标与目标位置的坐标归一化至[0, 999]范围。随后，将原始数据中离散的目标与地标描述性文本，整合到我们设计的结构化指令模板中。此举将原本孤立的文本描述，转化为一个明确的、具有时序和逻辑依赖关系的动态规划任务；同时将轨迹点和姿态角序列拆分为任务步骤，并用<|navigation|>标签记录，形成答案输出。通过该流程，每条轨迹均形成结构化的输入输出对，使模型的输出不再是单一的终点坐标，而是一个可执行的、分步展开的动作序列，从而模拟了从环境感知到路径生成的完整动态推理链。

### 3.5 任务5：定位描述

定位描述任务旨在生成图像描述的同时，将文本中的关键信息与图像中的特定区域进行显式对齐，即在自然语言叙述中体现目标的空间位置和外观属性。不同于仅关注全局语义的传统图像描述，该任务要求模型依据检测得到的定位信息生成与目标区域高度相关的语句，从而实现视觉内容与语言表达之间的精确映射与绑定。

#### 3.5.1 数据来源

该任务的数据来源于DOTA[25]和DIOR[20]两个目标检测数据集。其中，DOTA数据集收集自多种传感器和平台，共包含11 268幅图像、18类目标及约1 800 000个目标实例。DIOR数据集则包含23 463张图像和192 472个目标实例，覆盖20个目标类别，图像采集自Google Earth，空间分辨率范围为0.5～30 m。这两个数据集均覆盖多种常见场景，兼顾不同成像条件、天气状况以及季节特征差异，具有多样性。尽管上述数据集中提供了高质量的目标位置与类别标注，但其本身仅为检测任务设计，缺乏对图像内容的自然语言描述。为此，本文对其进行了关键性的任务扩展与语义增强。

#### 3.5.2 任务设置

在定位描述任务中，模型的输入为单幅遥感图像和一条自然语言指令，指令明确要求模型在生成描述时同时提供目标的空间位置。例如，在图5所示的样例中，输入指令用于引导模型在一句话中描述图像中的所有目标及其对应位置。模型的输出为结构化的定位描述，要求既包含自然语言描述，又提供目标类别与精确位置，实现视觉内容与文本描述的显式绑定。

#### 3.5.3 数据集制作流程

在数据集制作过程中，对于目标检测数据集中的图像， 首先将图像统一裁剪至标准尺寸，并剔除不含目标的图像样本。随后，基于原始目标检测标签整合每个目标的类别及相关信息。随后，在保持原有数据划分不变的前提下，本文为每幅图像构建了一批高质量的预训练样本。具体而言，首先基于开源数据集利用规则生成初始描述语言，并进一步为LLaVA-1.5[26]设计专用提示词，使模型能够生成更自然、细致且符合遥感语义的文本描述。为确保合成标注的可靠性，本文开发了自动化自评估与验证系统。系统通过计算生成文本与开源数据集中对应文本的语义相似度来评估描述质量；当相似度出现显著偏差时，将自动触发纠正流程，对描述进行修正并重新生成，保证标注的准确性与一致性。此外，本文还设计了模型的自我校验机制，通过比对生成描述与规则模板之间的类别及数量一致性，自动识别并纠正常见的语义冲突或遗漏，从而进一步提升预训练数据的质量与稳健性。最终构建了两个数据集DIOR-GC和DOTA-GC。其中，DIOR-GC包含训练集11 381条和测试集11 540条，总计22 921条样本；DOTA-GC包含训练集37 617条和测试集11 249条，总计48 866条样本。

#### 3.5.4 语言描述质量评估分析

表5汇总了DIOR-GC和DOTA-GC两个数据集在训练集与测试集的句子数量、单词数量及平均句

指令：Write all the objects in the picture and their position within one sentence.
答案：An empty parking lot with two <|ref|>basketball courts<|/ref|><|det|>[[158, 394, 302, 626], [400, 368, 544, 600]]<|/det|> in the middle and four <|ref|>vehicles<|/ref|><|det|>[[764, 367, 812, 387], [774, 420, 824, 445], [774, 533, 830, 558], [904, 410, 952, 434]]<|/det|> surrounding them.

图 5 定位描述任务问答对示例

表 5　定位描述数据集信息统计表

| 训练集 | 句子数量 | 单词数量 | 平均句子长度 |
|---|---|---|---|
| DIOR-GC训练集 | 11381 | 151804 | 13.34 |
| DIOR-GC测试集 | 11540 | 164248 | 14.23 |
| DOTA-GC训练集 | 37617 | 932049 | 24.78 |
| DOTA-GC测试集 | 11249 | 239331 | 21.28 |
| 总数 | 71787 | 1487432 | 18.41 |

子长度的统计信息。从中可以看出，DOTA-GC的句子相对较长(训练集平均24.78个单词，测试集21.28个单词)，而DIOR-GC的句子较为简短(训练集平均13.34个单词，测试集14.23个单词)。总体而言，整个定位描述数据集覆盖71 787条句子、1 487 432个单词，平均每句约18.41个单词。这表明数据集既包含简洁描述，也涵盖较长、信息丰富的句子，能够为视觉语言模型提供多样化的训练语料，有助于提升模型在生成定位描述时的准确性和表达能力。

### 3.6　任务6：多模态感知任务

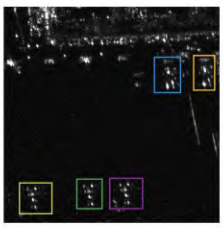
多模态感知任务涵盖目标检测、图像分类、图像描述、目标计数4个子任务，每个子任务均提供光学、SAR和红外3种遥感数据模态作为输入。该任务旨在评估模型在不同遥感模态下的感知能力，基于用户指令，要求模型能够根据单一模态的数据完成感知任务。通过对多种模态的独立使用，能够充分考察模型在应对不同传感器特性、成像条件和噪声特征时的表现，为多模态遥感数据的分析提供全面的基准和评估参考。

#### 3.6.1　数据来源

本任务使用了光学、SAR和红外3种数据源。由于当前用于图文任务的 SAR 和红外数据极其有限，本文基于现有的检测数据集生成了SAR和红外的图文对样本，以扩展任务所需数据。具体来说，SAR数据来自SARDet-100K数据集[21]，该数据集整合了来自10个公开高质量SAR数据集的图像，共包含116 598张图像和245 653个目标实例，共有6个目标类别。红外数据则由多个开源数据集[27–32]整合而成，主要覆盖海上、街景等场景。相较之下，光学数据由于资源丰富，检测任务直接采用DIOR[20]数据集并调整格式，分类任务选取AID[22]与NWPU-RESISC45[23]数据集并调整格式，计数和图像描述任务则基于DIOR[20]和DOTA[25]数据集生成。

#### 3.6.2　任务设置

本研究的多模态感知任务涵盖了目标检测、图像分类、图像描述和目标计数4类子任务，分别针对SAR、红外和光学3种模态数据进行设置，红外和SAR数据的问答对可视化如图6所示。对于检测任务，指令要求模型在图像中定位目标并给出其精确坐标，其中SAR数据侧重识别如airplanes等大类目标，红外数据则需检测包括liners和warships等细粒度类别。分类任务的指令为根据输入图像输出对应的场景类别，以评估模型对整体场景语义的识别能力。描述任务要求模型根据图像生成简洁的自然语言描述，SAR数据的描述侧重目标数量与形态，红外数据的描述则需明确区分细粒度类别，从而综合考察模型在多模态条件下的视觉理解与语言生成能力。在计数任务中，模型的输入为一幅图像及自然语言提问，输出为对应的目标数量或计数结果。



检测指令：Find the objects and provide the coordinates of each object:
答案：There are 5 <|ref|>airplanes<|/ref|><|det|>[[72, 813, 226, 959], [489, 792, 645, 936], [337, 796, 458, 936], [878, 224, 979, 388], [696, 232, 821, 399]]<|/det|> in the image.
分类指令：Output the scene categories corresponding to the given image:
答案：Aircraft.
描述指令：Share a concise interpretation of the image provided:
答案：The image contains 5 airplanes with a medium shape.
计数指令：What is the total number of airplanes visible in this scene?
答案：5

(a) SAR数据集在检测、分类、描述、计数任务的可视化

检测指令：Locate objects and indicate their precise positions in the images, requiring detection of fine-grained categories such as truck, liners, and warships:
答案：There are 1 <|ref|>fishing boat<|/ref|><|det|>[[0, 500, 460, 590]]<|/det|> and 1 <|ref|>warship<|/ref|><|det|>[[560, 460, 660, 490]]<|/det|> in the image.
分类指令：Output the scene categories corresponding to the given image:
答案：Ship.
描述指令：Offer a succinct explanation of the picture presented, including fine-grained categories such as such as truck, liners, and warships:
答案：One fishing boat and one warship are visible in the image.
计数指令：What is the total number of fishing boats visible in this scene?
答案：1

(b) 红外数据集在检测、分类、描述、计数任务的可视化

图 6 多模态感知任务问答对示例

#### 3.6.3 数据集制作流程

首先，本文对SARDet-100K检测数据集的原始格式进行调整，将其转换为适用于图文模型的目标检测数据，并按9:1比例划分为训练集(104 985张)和测试集(11 613张)；随后，本文采用3.5节数据生成的方法，选择Llama3.1[33]模型并为其设计专用提示词，基于同样严格的约束并最终生成得到数据集SAR-CAP。每张图像生成5条不同的文本描述，数据划分比例仍保持9:1；另外，本文将数据转换为分类任务格式的SAR-CLA，并采用1:9的比例划分，训练集11 612条图文对，测试集104 985条图文对，其中分类标签直接对应于场景中包含的目标类别名称；最后，根据检测数据的原始标注文件，提取其中的目标类别与对应的边界框信息，并通过解析坐标数量确定每类目标在图像中的实例数，利用预定义的英文问题模板随机生成自然语言指令，并将目标类别名称根据数量规则转化计数任务数据集SAR-COUNT，训练集与测试集文件分别有106 098条和11 705条，数据划分比例为9:1。

基于6个开源红外数据集，本文沿用与SAR数据相同的数据制作流程。不同于SAR数据更关注大类的特性，红外数据同时注重粗粒度与细粒度目标的识别。因此，本文在红外数据的指令设计中增加了粒度指向性的提示语，例如“only requiring detection of coarse categories”或“requiring detection of fine-grained categories”。在检测任务中，本文构建了名为IR-DET的红外检测图文数据集，包含训练集50 711对图文对和测试集5 642对图文对，比例为9:1；对于描述任务，本文构建了IR-CAP，其样本比例与IR-DET 保持一致；由于红外图像内容较为复杂，分类任务中本文仅选取街景和船只两类构建二分类数据集，命名为IR-CLA；最后，根据计数问题模板，生成数据集IR-COUNT，其中训练数据96 741条和测试数据10 824条。

最后，对于光学数据，检测和分类任务直接采用开源数据集，仅对指令格式进行了微调以适配模型输入。对于图像描述任务，本文同样采用生成扩充的方法，其具体流程与定位描述章节中介绍的生成方法一致，最终构建得到DIOR-CAP和DOTA-CAP数据集。其中，DIOR-CAP包含11 725张训练图像，通过3种不同的提问方式生成35 175条训练数据；测试集包含11 540条样本，经扩展为5句话形式，共得到57 700条测试数据，总计92 875条。同样地，DOTA-CAP 则包含250 905条训练数据，测试集包含11 249条样本，经扩展为5句话形式，共得到56 245条测试数据，总计307 150条。二者合计共包含400 025条数据。通过同样的计数数据集扩展方式，得到了DIOR-COUNT数据集，其中训练集16 478条，测试集18 726条，总计35 204条数据，以及DOTA-COUNT数据集，其中训练集60 403条，测试集18 029条，总计78 432条数据。

#### 3.6.4 语言描述质量评估分析

表6和表7分别统计了多模态感知任务中图像描述和分类数据集的基本信息。首先，对同一图像的多条指令进行去重处理，仅保留其中1条有效指令，得到表6。由表中可见在SAR-CAP数据集包含训练集524 925条句子和测试集58 065条句子，平均每条句子长度约为9.46个单词；而IR-CAP数据集包含训练集132 410条句子和测试集14 735条句子，平均每条句子长度约10.08个单词。整体来看，红外数据的描述句子略长于SAR数据，表明红外图像的描述任务在语言表达上更为丰富，这主要是由于红外图像中的目标更多样和复杂。另外，DIOR-CAP数据集的训练集包含11 725条句子，测试集包含57 700条句子，DOTA-CAP数据集的训练集包含83 635条句子，测试集包含56 245条句子，平均长度为15.84个单词。

从表7的分类数据统计来看，SAR分类数据集覆盖多个目标组合类别，红外分类数据集仅包含街景和船只两类。对比两种模态数据可以发现，SAR数据在类别组合和多目标场景上更为复杂，而红外数据则相对简洁集中。

## 4 关联评价指标和基线结果

### 4.1 不同任务的评价指标介绍

关系推理及关系检测任务：采用F1分数(F1-score)衡量预测关系与真实关系的一致性。

指令分解任务：关系部分计算F1-score，目标

表 6 多模态感知中图像描述数据集统计

| 数据类型 | 训练集/测试集 | 句子数量 | 单词数量 | 平均句子长度 |
|---|---|---|---|---|
| SAR | SAR-CAP训练集 | 524925 | 4950300 | 9.43 |
| | SAR-CAP测试集 | 58065 | 550216 | 9.48 |
| | 总数 | 582990 | 5500516 | 9.46 |
| 红外 | IR-CAP训练集 | 132410 | 1330173 | 10.05 |
| | IR-CAP测试集 | 14735 | 148973 | 10.11 |
| | 总数 | 147145 | 1479146 | 10.08 |
| 可见光 | DIOR-CAP训练集 | 11725 | 157232 | 13.41 |
| | DIOR-CAP测试集 | 57700 | 738907 | 12.81 |
| | DOTA-CAP训练集 | 83635 | 1486176 | 17.77 |
| | DOTA-CAP测试集 | 56245 | 1090289 | 19.38 |
| | 总数 | 209305 | 3472604 | 15.84 |

部分计算平均精度均值(mean Average Precision at IoU=0.5, mAP50)。

任务调度任务：根据最终生成轨迹的正确性，使用导航误差(Navigation Error, NE)、成功率(Success Rate, SR)、理想成功率(Oracle Success Rate, OSR)以及路径长度加权成功率(Success weighted by Path Length, SPL)4个指标进行评估。其中NE以米(m)为单位，其余指标均以百分比(%)表示。

定位描述任务：在图像描述评价指标(BLEU-4, CIDEr)的基础上，本文额外引入mAP50，用于衡量生成文本与目标框之间的定位精度。在评估过程中，图像描述指标的计算采用5句扩展方式，并在对应的JSON文件中提供了相关内容。

多模态感知任务：目标检测子任务计算mAP50；分类子任务计算分类准确率；图像描述子任务计算BLEU-4, METEOR和ROUGE-L；目标计数任务计算预测精度。

**表 7　多模态感知中图像分类数据集统计**

| 数据类型 | 类别名称 | 训练集 | 测试集 | 训练/测试比例 |
|---|---|---|---|---|
| SAR | Aircraft | 3037 | 16835 | 0.1804 |
| | Aircraft and tank | 3 | 4 | 0.7500 |
| | Bridge | 1697 | 16168 | 0.1050 |
| | Bridge and harbor | 16 | 131 | 0.1221 |
| | Bridge and ship | 13 | 230 | 0.0565 |
| | Bridge and tank | 33 | 329 | 0.1003 |
| | Bridge, harbor and tank | 3 | 25 | 0.1200 |
| | Car | 103 | 941 | 0.1095 |
| | Harbor | 138 | 1255 | 0.1100 |
| | Harbor and tank | 9 | 82 | 0.1098 |
| | Ship | 6470 | 67211 | 0.0963 |
| | Ship and tank | 13 | 287 | 0.0453 |
| | Tank | 77 | 1487 | 0.0518 |
| | 总数 | 11612 | 104985 | 0.1505 |
| 红外 | Ship | 1593 | 14354 | 0.1110 |
| | Street | 4036 | 36370 | 0.1110 |
| | 总数 | 5629 | 50724 | 0.1110 |

### 4.2 基线方法结果

基于团队前期工作的实验结果[34,35]，本文提供了两种经过遥感数据专门预训练与微调的领域模型(RingMo-Agent[34], RingMoGPT[35])，以及两种在通用多模态数据上预训练的模型(MiniGPT-v2[7], DeepSeek-VL2[36])。对于后者，除在ReCon1M-DEC数据集上进行微调外，其余任务均采用零样本泛化测试。各项任务的详细评估结果可为未来研究提供明确的性能参照。

#### 4.2.1 关系推理任务

从表8结果可见，通用视觉语言模型在该任务中的表现整体较低，说明其对遥感领域关系推理任务的适应性和理解能力不足。

#### 4.2.2 指令分解任务

在该分解任务中(见表9)，遥感视觉语言模型RingMo-Agent在mAP50 (24.20%)和F1-Score (32.85%)上均显著优于通用视觉语言模型(最高分别为19.80%和15.19%)，表明其在同时处理关系识别与目标定位方面具备更强的适应性。

#### 4.2.3 任务调度任务

如表10所示，在CityNav数据集的任务调度任务中，专业模型AerialVLN在所有数据划分上均取得最低的导航误差与最高的成功率，表现最优；通用遥感视觉语言模型RingMo-Agent尽管在NE与SR上落后于AerialVLN，但相比专业模型Seq2Seq与CMA仍具有明显优势。

**表 8　关系推理任务在ReCon1M-REL数据集上的结果(%)**

| 模型类型 | 模型方法 | F1-Score ↑ |
|---|---|---|
| 通用视觉语言模型(零样本评估) | MiniGPT-v2[7] | 0.00 |
| | DeepSeek-VL2[36] | 0.30 |
| 遥感视觉语言模型(微调后评估) | RingMo-Agent[34] | 90.23 |

**表 9　指令分解任务在ReCon1M-DEC数据集上的结果(%)**

| 模型类型 | 模型方法 | mAP50 ↑ | F1-Score ↑ |
|---|---|---|---|
| 通用视觉语言模型(微调后评估) | MiniGPT-v2[7] | 11.50 | 15.19 |
| | DeepSeek-VL2[36] | 19.80 | 10.32 |
| 遥感视觉语言模型(微调后评估) | RingMo-Agent[34] | 24.20 | 32.85 |

**表 10　任务调度任务在CityNav数据集上的结果**

| 模型类型 | 模型方法 | 测试集 | | | |
|---|---|---|---|---|---|
| | | NE ↓ | SR ↑ | OSR ↑ | SPL ↑ |
| 专业模型(微调后评估) | Seq2Seq[37] | 245.30 | 1.50 | 8.34 | 1.30 |
| | CMA[38] | 252.60 | 0.82 | 9.70 | 0.79 |
| | AerialVLN + GSM[19] | 85.10 | 6.72 | 18.21 | 5.16 |
| 遥感视觉语言模型(微调后评估) | RingMo-Agent[34] | 149.60 | 4.74 | 18.94 | 4.17 |

#### 4.2.4 定位描述任务

如表11所示，相比基线版本，RingMoGPT在各项语言生成指标上均有小幅提升，其中CIDEr提升最明显(+1.4)，说明生成文本的相关性和细节表达更优；在视觉检测任务上提升更为显著，DIOR-GC和DOTA-GC的mAP50分别提升了20.4%和10.9%，表明优化后的模型在跨任务迁移与多模态理解方面具备更强的综合能力。

#### 4.2.5 多模态感知任务

在多模态感知任务的图像描述任务中，遥感视觉语言模型RingMo-Agent在SAR-CAP和IR-CAP数据集上均显著优于通用视觉语言模型(见表12)。RingMo-Agent的BLEU系列指标、METEOR和ROUGE-L均远高于零样本评估的MiniGPT-v2和DeepSeek-VL2，显示其在生成与图像内容高度相关的描述方面具有更强的语言理解和多模态融合能力，同时体现出针对遥感图像的优良适应性。

如表13所示，在将DIOR-CAP和DOTA-CAP数据集作为整体进行多模态感知任务评估时，RingMoGPT相较于其基线模型在各项指标均有提升：METEOR从28.8提升至30.4，ROUGE-L从55.4提升至60.1，CIDEr从72.5提升至99.2，整体生成质量和相关性得分均有所提高。

如表14所示，在IR-DET数据集的多模态感知目标检测任务中，遥感视觉语言模型RingMo-Agent显著优于所有通用视觉语言模型，mAP50达到59.88%，而通用模型在不明确检测类名的情况下难以获得有效结果。

如表15所示，在多模态感知任务的分类评估中，RingMo-Agent在SAR-CLA数据集的分类准确率为92.67%，在IR-CLA数据集的分类准确率为99.45%。相比之下，在不给予类名提示的情况下，MiniGPT-v2在SAR-CLA和IR-CLA的分类准确率分别为5.87%和60.52%，DeepSeek-VL2分别为4.40%和45.15%。这些数据反映了不同模型在不同遥感模态上的分类性能差异。

**表 11 定位描述任务在DIOR-GC和DOTA-GC数据集上的结果(%)**

| 模型类型 | 模型方法 | BLEU-1 ↑ | BLEU-2 ↑ | BLEU-3 ↑ | BLEU-4 ↑ | METEOR ↑ | ROUGE-L ↑ | CIDEr ↑ | mAP50 (DIOR-GC) ↑ | mAP50 (DOTA-GC) ↑ |
|---|---|---|---|---|---|---|---|---|---|---|
| 遥感视觉语言模型(微调后评估) | RingMoGPT[35]基线 | 67.5 | 54.2 | 43.6 | 34.8 | 28.7 | 58.6 | 92.7 | 44.9 | 35.6 |
| | RingMoGPT[35] | 68.5 | 55.1 | 44.2 | 35.7 | 29.4 | 57.3 | 94.1 | 65.3 | 46.5 |

**表 12 多模态感知任务在SAR-CAP和IR-CAP数据集上的结果(%)**

| 模型类型 | 模型方法 | SAR-CAP | | | | | | IR-CAP | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | BLEU-1 ↑ | BLEU-2 ↑ | BLEU-3 ↑ | BLEU-4 ↑ | METEOR ↑ | ROUGE-L ↑ | BLEU-1 ↑ | BLEU-2 ↑ | BLEU-3 ↑ | BLEU-4 ↑ | METEOR ↑ | ROUGE-L ↑ |
| 通用视觉语言模型(零样本评估) | MiniGPT-v2[7] | 7.00 | 3.64 | 1.59 | 0.60 | 7.67 | 9.25 | 5.65 | 3.35 | 1.92 | 0.98 | 7.62 | 8.67 |
| | DeepSeek-VL2[36] | 12.52 | 5.88 | 1.88 | 0.60 | 10.65 | 14.10 | 13.95 | 7.57 | 3.47 | 1.43 | 12.48 | 15.12 |
| 遥感视觉语言模型(微调后评 | RingMo-Agent[34] | 55.93 | 44.49 | 33.57 | 23.94 | 25.06 | 51.12 | 56.84 | 40.45 | 29.17 | 21.50 | 26.15 | 43.13 |

**表 13 多模态感知任务在DIOR-CAP和DOTA-CAP数据集上的结果(%)**

| 模型类型 | 模型方法 | METEOR ↑ | ROUGE-L ↑ | CIDEr ↑ |
|---|---|---|---|---|
| 遥感视觉语言模型(微调后评估) | RingMoGPT[35]基线 | 28.80 | 55.40 | 72.50 |
| | RingMoGPT[35] | 30.40 | 60.10 | 99.20 |

**表 14 多模态感知任务在IR-DET数据集上的结果(%)**

| 模型类型 | 模型方法 | mAP50 ↑ |
|---|---|---|
| 通用视觉语言模型(零样本评估) | MiniGPT-v2[7] | 0 |
| | DeepSeek-VL2[36] | 0 |
| 遥感视觉语言模型(微调后评估) | RingMo-Agent[34] | 59.88 |

表 15　多模态感知任务在SAR-CLA和IR-CLA数据集上的结果(%)

| 模型类型 | 模型方法 | SAR-CLA分类准确率 ↑ | IR-CLA分类准确率 ↑ |
|---|---|---|---|
| 通用视觉语言模型(零样本评估) | MiniGPT-v2[7] | 5.87 | 60.52 |
| | DeepSeek-VL2[36] | 4.40 | 45.15 |
| 遥感视觉语言模型(微调后评估) | RingMo-Agent[34] | 92.67 | 99.45 |

## 5　结论与展望

本文面向遥感智能基础模型的感知——认知建模需求，构建并公开了一个大规模遥感图文指令数据集，涵盖9类关键任务、21个子数据集，覆盖光学、SAR和红外等多模态数据，累计超过2 000 000样本。通过统一的图文指令范式与结构化输入输出格式，本文数据集不仅覆盖了目标检测、图像描述、分类和计数等静态感知任务，还进一步拓展至关系推理、关系检测、任务调度和指令分解等高层认知任务，具备良好的通用性与拓展性。在遥感视觉语言模型上的基线实验表明，构建的数据集能有效支持模型在多模态理解、多任务执行与跨模态生成等任务的训练。随着遥感影像的时空分辨率持续提升和多源感知手段的不断丰富，遥感智能体将逐步从被动识别走向主动规划与自适应交互。构建更加开放、动态、交互式的数据体系将成为推动基础模型进一步发展的关键方向。后续工作中，本文计划进一步扩展该数据集的时序性与对话性，探索多轮交互、多智能体协同等复杂任务场景，联合推动多模态遥感任务的统一评测平台建设，促进遥感智能体能力的标准化量化评估。

## 参 考 文 献


[1] 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.

[2] HU Huiyang, WANG Peijin, BI Hanbo, *et al*. RS-vHeat: Heat conduction guided efficient remote sensing foundation model[C]. The IEEE/CVF International Conference on Computer Vision, Honolulu, The United States of America, 2025: 9876–9887.

[3] CHANG Hao, WANG Peijin, DIAO Wenhui, *et al*. Remote sensing change detection with bitemporal and differential feature interactive perception[J]. *IEEE Transactions on Image Processing*, 2024, 33: 4543–4555. doi: 10.1109/TIP.2024.3424335.

[4] SHI Qian, HE Da, LIU Zhengyu, *et al*. Globe230k: A benchmark dense-pixel annotation dataset for global land cover mapping[J]. *Journal of Remote Sensing*, 2023, 3: 0078. doi: 10.34133/remotesensing.0078.

[5] HU Fengming, XU Feng, WANG R, *et al*. Conceptual study and performance analysis of tandem multi-antenna spaceborne SAR interferometry[J]. *Journal of Remote Sensing*, 2024, 4: 0137. doi: 10.34133/remotesensing.0137.

[6] MEI Shaohui, LIAN Jiawei, WANG Xiaofei, *et al*. A comprehensive study on the robustness of deep learning-based image classification and object detection in remote sensing: Surveying and benchmarking[J]. *Journal of Remote Sensing*, 2024, 4: 0219. doi: 10.34133/remotesensing.0219.

[7] CHEN Jun, ZHU Deyao, SHEN Xiaoqian, *et al*. MiniGPT-v2: Large language model as a unified interface for vision-language multi-task learning[J]. arXiv: 2310.09478, 2023. doi: 10.48550/arXiv.2310.09478.

[8] PENG Zhiliang, WANG Wenhui, DONG Li, *et al*. Kosmos-2: Grounding multimodal large language models to the world[J]. arXiv: 2306.14824, 2023. doi: 10.48550/arXiv.2306.14824.

[9] CHEN Keqin, ZHANG Zhao, ZENG Weili, *et al*. Shikra: Unleashing multimodal LLM's referential dialogue magic[J]. arXiv: 2306.15195, 2023. doi: 10.48550/arXiv.2306.15195.

[10] 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.

[11] 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.

[12] 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.

[13] 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.

[14] 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.

[15] 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.

[16] 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.

[17] 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: 2406.10100, 2024. doi: 10.48550/arXiv.2406.10100.

[18] LI Xiang, DING Jian, and MOHAMED E. 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.

[19] LEE J, MIYANISHI T, KURITA S, *et al*. CityNav: A large-scale dataset for real-world aerial navigation[C]. The IEEE/CVF International Conference on Computer Vision, Honolulu, The United States of America, 2025: 5912–5922.

[20] 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.

[21] LI Yuxuan, LI Xiang, LI Weijie, *et al*. SARDet-100K: Towards open-source benchmark and toolkit for large-scale SAR object detection[C]. The 38th International Conference on Neural Information Processing Systems, Vancouver, Canada, 2024: 4079.

[22] 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.

[23] 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.

[24] YAN Qiwei, DENG Chubo, LIU Chenglong, *et al*. ReCon1M: A large-scale benchmark dataset for relation comprehension in remote sensing imagery[J]. *IEEE Transactions on Geoscience and Remote Sensing*, 2025, 63: 4507022. doi: 10.1109/TGRS.2025.3589986.

[25] 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.

[26] 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.

[27] SUO Jiashun, WANG Tianyi, ZHANG Xingzhou, *et al*. HIT-UAV: A high-altitude infrared thermal dataset for Unmanned Aerial Vehicle-based object detection[J]. *Scientific Data*, 2023, 10(1): 227. doi: 10.1038/s41597-023-02066-6.

[28] 李春柳, 王水根. 红外海上船舶数据集[EB/OL]. https://openai.raytrontek.com/apply/Sea_shipping.html, 2021.
LI Chunliu and WANG Shuigen. Infrared sea-shipping dataset[EB/OL]. https://openai.raytrontek.com/apply/Sea_shipping.html, 2021.

[29] 刘晴, 徐召飞, 金荣璐, 等. 红外安防数据库[EB/OL]. https://openai.raytrontek.com/apply/Infrared_security.html, 2021.
LIU Qing, XU Zhaofei, JIN Ronglu, *et al*. Infrared-security dataset[EB/OL]. https://openai.raytrontek.com/apply/Infrared_security.html, 2021.

[30] 刘晴, 徐召飞, 王水根. 红外航拍人车检测数据集[EB/OL]. http://openai.raytrontek.com/apply/Aerial_mancar.html, 2021.
LIU Qing, XU Zhaofei, and WANG Shuigen. Infrared aerial-mancar dataset[EB/OL]. http://openai.raytrontek.com/apply/Aerial_mancar.html, 2021.

[31] 李钢强, 王建生, 王水根. 双光车载场景数据库[EB/OL]. http://openai.raytrontek.com/apply/Double_light_vehicle.html, 2021.
LI Gangqiang, WANG Jiansheng, and WANG Shuigen. Double-light-vehicle dataset[EB/OL]. http://openai.raytrontek.com/apply/Double_light_vehicle.html, 2021.

[32] 山东大学光学高等研究中心. 远海(10–12km)船舶的目标检测数据集[EB/OL]. http://www.core.sdu.edu.cn/info/1133/2174.htm, 2020.
Center for Optics Research and Engineering of Shandong University. Open-sea (10–12 km) ship object detection dataset[EB/OL]. http://www.core.sdu.edu.cn/info/1133/2174.htm, 2020.

[33] GRATTAFIORI A, DUBEY A, JAUHRI A, *et al*. The llama 3 herd of models[J]. arXiv: 2407.21783, 2024. doi: 10.48550/arXiv.2407.21783.

[34] HU Huiyang, WANG Peijin, FENG Yingchao, *et al*. RingMo-Agent: A unified remote sensing foundation model for multi-platform and multi-modal reasoning[J]. arXiv: 2507.20776, 2025. doi: 10.48550/arXiv.2507.20776.

[35] 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.

[36] WU Zhiyu, CHEN Xiaokang, PAN Zizheng, *et al*. Deepseek-vl2: Mixture-of-experts vision-language models for advanced multimodal understanding[J]. arXiv: 2412.10302, 2024. doi: 10.48550/arXiv.2412.10302.

[37] ANDERSON P, WU Q, TENEY D, *et al*. Vision-and-language navigation: Interpreting visually-grounded navigation instructions in real environments[C]. The IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, USA, 2018: 3674–3683. doi: 10.1109/CVPR.2018.00387.

[38] LIU Shuobo, ZHANG Hongsheng, QI Yuankai, *et al*. AeriaLVLN: Vision-and-language navigation for UAVs[C]. The IEEE/CVF International Conference on Computer Vision, Paris, France, 2023: 15338–15348. doi: 10.1109/ICCV51070.2023.01411.

王佩瑾：女，助理研究员，研究方向为遥感图像智能解译.
胡会扬：女，博士生，研究方向为遥感图像智能解译.
冯瑛超：男，助理研究员，研究方向为遥感图像智能解译.
刁文辉：男，副研究员，研究方向为遥感图像智能解译.
孙　显：男，研究员，研究方向为计算机视觉与遥感图像理解.

责任编辑：余　蓉

# A Large-Scale Multimodal Instruction Dataset for Remote Sensing Agents

WANG Peijin[①②③④]　HU Huiyang[①②③④]　FENG Yingchao[①④]
DIAO Wenhui[①②③④]　SUN Xian[①②③④]

[①](*Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing* 100190, *China*)
[②](*University of Chinese Academy of Sciences, Beijing* 100190, *China*)
[③](*School of Electronic, Electrical and Communication Engineering, University of Chinese Academy of Sciences, Beijing* 100049, *China*)
[④](*Key Laboratory of Target Cognition and Application Technology*(*TCAT*), *Beijing* 100190, *China*)

**Abstract**:

**Objective**　The rapid advancement of Remote Sensing (RS) technology has reshaped Earth observation research, shifting the field from static image analysis to intelligent, goal-oriented cognitive decision-making. Modern RS systems are expected to perceive complex scenes, reason over heterogeneous information, decompose high-level objectives into executable subtasks, and make decisions under uncertainty. These requirements motivate the development of RS agents, which extend perception models to include reasoning, planning, and interaction functions. However, existing RS datasets remain task-centric and fragmented, as they are usually designed for single-purpose supervised learning such as object detection or land-cover classification. They seldom support multimodal reasoning, instruction following, or multi-step decision-making, all of which are essential for agentic workflows. Current RS vision-language datasets also have limited scale, constrained modality coverage, and simplified text annotations, with insufficient use of non-optical data such as Synthetic Aperture Radar (SAR) and infrared imagery. They further lack instruction-driven interactions that reflect real human-agent collaboration. This study constructs a large-scale multimodal image-text instruction dataset tailored for RS agents. The objective is to establish a unified data foundation that supports perception, reasoning, planning, and decision-making. By training models on structured instructions across diverse modalities and task categories, the dataset supports the development and evaluation of next-generation RS foundation models with agentic capability.

**Methods**　The dataset is built through a systematic and extensible framework that integrates multi-source RS imagery with instruction-oriented textual supervision. A unified input-output paradigm is defined to ensure compatibility across heterogeneous tasks and model architectures. This paradigm formalizes interactions between visual inputs and language instructions, allowing models to process image pixels, text descriptions, spatial coordinates, region references, and action-oriented outputs. A standardized instruction schema encodes

task objectives, constraints, and expected responses in a consistent format. The construction process includes three stages. (1) Data collection and integration: multimodal RS imagery is aggregated from authoritative sources, covering optical, SAR, and infrared modalities with different spatial resolutions, scene types, and geographic distributions. (2) Instruction generation: a hybrid strategy combines rule-based templates with refinement by Large Language Models (LLMs). Template-based generation ensures task completeness and structural consistency, whereas LLM rewriting improves linguistic diversity and instruction complexity. (3) Task categorization and organization: the dataset is organized into nine core task categories and 21 sub-datasets that span low-level perception, mid-level reasoning, and high-level decision-making. A validation pipeline performs automated syntax and format checks, cross-modal consistency verification, and manual review of representative samples to ensure semantic alignment between images and instructions.

**Results and Discussions** The dataset contains more than 2 million multimodal instruction samples, making it one of the largest and most comprehensive instruction resources in the RS domain. The inclusion of optical, SAR, and infrared imagery supports cross-modal learning and reasoning across heterogeneous sensing mechanisms. Compared with existing RS datasets, this dataset emphasizes instruction diversity, task compositionality, and agent-oriented interaction rather than isolated perception tasks. Baseline experiments conducted using state-of-the-art multimodal LLMs and RS foundation models show that the dataset supports evaluation across the full spectrum of agentic capabilities, from visual grounding and reasoning to high-level decision-making. The experiments also highlight challenges inherent to RS data, including extreme scale variation, dense object distributions, and long-range spatial dependencies. These challenges indicate important research directions for improving multimodal reasoning and planning in complex RS environments.

**Conclusions** This work presents a large-scale multimodal image-text instruction dataset designed for RS agents. By organizing data across nine task categories and 21 sub-datasets, it provides a unified and extensible benchmark for agent-centric RS research. The contributions include: (1) a unified multimodal instruction paradigm for RS agents; (2) a 2-million-sample dataset covering optical, SAR, and infrared modalities; (3) empirical validation demonstrating support for end-to-end agentic workflows from perception to decision-making; and (4) a comprehensive evaluation benchmark based on baseline experiments. Future work will extend the dataset to temporal and video-based RS scenarios, integrate dynamic decision-making processes, and further improve reasoning and planning capability in real-world, time-varying environments.

**Key words**: Remote sensing foundation models; Multimodal instruction datasets; Perception-cognition-decision