arXiv:2512.17319v1 [cs.CV] 19 Dec 2025

超高分辨率遥感多模态大语言模型基准

Yunkai Dang*、Meiyi Zhu*、Donghao Wang[1],Yizhuo Zhang[1],
Jiacheng Yang[1],Qi Fan[1], Yuekun Yang[1],Wenbin Li†、Feng Miao[2],
Yang Gao[1]人工智能科学与技术学院，南京大学[2]物理学院，南京大学
yunkaidang@smail.nju.edu.cn, liwenbin@nju.edu.cn

摘要

多模态大语言模型（多模态大语言模型）在现有遥感（*RS*）基准数据集上展现出强大的感知和推理能力。然而，这些基准数据集大多依赖低分辨率遥感图像，而少数高分辨率基准数据集的推理任务设计存在缺陷。我们发现，在不查看图像的情况下，纯文本大语言模型在遥感推理任务上的表现也能媲美多模态视觉语言模型。这揭示了当前基准数据集与其旨在评估的视觉理解能力之间存在严重错位。为实现忠实评估，我们推出 *RSHR-Bench*，这是一个用于遥感视觉理解与推理的超高分辨率基准。其中包含长边至少为 4000 像素的 *5,329* 张全景图像。每张图像最多包含 $\sim 3 \times 10^8$ 像素，图像来源于广泛使用的遥感语料库和无人机数据集。我们设计了四类任务：多项选择视觉问答、开放式视觉问答、图像描述和单图评估。这些任务涵盖九种感知类别和四种推理类型，并支持多轮和多图像对话。为减少对语言先验的依赖，我们采用强大的大语言模型进行对抗式过滤，随后进行严格的人工验证。总计，我们构建了 *3,864* 项视觉问答任务（包括封闭集和开放式设置）、*3,913* 项图像描述任务，以及 *500* 对完全由人工撰写并验证的单图评估视觉问答对。在 *RSHR-Bench* 上评估一系列广泛的开源、闭源及遥感专用视觉语言模型后，我们发现它们在高分辨率场景中仍存在显著的性能差距。我们的代码已公开发布于 https://github.com/Yunkaidang/RSHR。

1. 引言

近年来，多模态大语言模型（MLLMs）的视觉理解与推理能力取得了显著进展 [1, 2, 5, 11, 26, 47, 69, 75] 。受高分辨率视频和自动驾驶等应用的推动

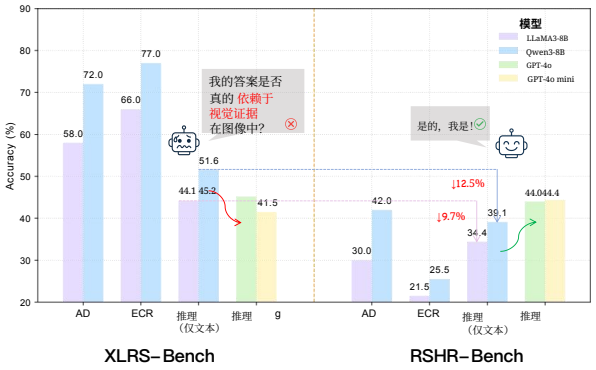


图 1. XLRS-Bench [63] 和 RSHR-Bench 上的准确率。任务：AD（异常检测）和 ECR（存在与计数推理）。我们报告两种输入设置下的平均推理准确率：纯文本 （Llama3-8B 和 Qwen3-8B）和多模态 （图像+文本；GPT-4o 和 GPT-4o mini）。RSHR-Bench 在纯文本和多模态设置之间表现出更大的差距，表明其对视觉信息的依赖更强。

，另一条研究路线致力于扩展 MLLMs [19, 32, 35, 53, 87] 处理高分辨率输入的能力，并出现了多个基准数据集 [66, 70, 86] ，将 4K/8K图像纳入评测这些模型的范围。相比之下，遥感通常涉及卫星和无人机在大范围地理区域内拍摄的超高分辨率图像。这些场景呈现极端的多尺度变化，且杂乱背景中分布着数量稀少的小型物体 [6]。然而，尽管高分辨率输入方面取得了进展，通用模型在实际遥感图像上的评测仍不充分，尤其是在原生空间分辨率和地理尺度下。因此，遥感领域的专用多模态大语言模型 [27, 49, 55, 62] 通常使用遥感数据进行微调，并采用基于图块的处理或直接调整尺寸的方法，以应对各种任务（感知、推理、图像描述）。

一系列遥感基准数据集 [13, 31, 46,

*共同贡献。
†通讯作者。

63, 64] 支持在多种遥感场景下进行评估。然而，这些研究仍侧重于低分辨率设置，依赖会模糊场景级上下文的小图块：VRSBench [31]主要使用 512×512切片；RSVQA [38] 采用 499×499；HRVQA [28]达到 1024×1024。近年来，更高分辨率的多模态基准数据集相继出现，规模提升至 8500×8500(XLRS-Bench [63]) 和 7099×6329 (LRS-VQA [43])。然而，推理任务设计仍存在局限：即使 没有视觉输入，XLRS-Bench 也能取得较高准确率；而 LRS-VQA 主要侧重于感知，而非推理。如图1所示，纯文本大语言模型在 XLRS-Bench 的存在性和计数推理任务上取得了77% 的准确率。此外，在所有推理任务中，纯文本大语言模型（Qwen3-8B）的准确率达到 51。6% ，超过了准确率为 45。2% 的 GPT-4o（多模态设置）。这些结果表明，模型可能通过利用文本线索和已有的世界知识来回答许多问题，而非忠实地解读视觉内容。

因此，对高分辨率遥感进行评估需要经过审慎的重新设计。我们强调四项关键挑战：**(1) 基准分辨率**。现有大多数基准数据集都在小图像上评估模型，而现实场景中的遥感图像要大得多——例如，单张 DOTA [71] 图像最多可包含 $\sim 2\times10^8$ 像素，PANDA [8] 则达到千兆像素级分辨率（即每帧 $\sim 2\times2.5^8$ 像素）。**(2) 高分辨率下的长程结构**。常见的切片流程使用的输入尺寸从 512 到 2,000 像素不等，这会割裂全局布局和目标群集。因此，它们很少能直接评估模型对整幅图像的理解能力。**(3) 推理任务设计未能控制大语言模型的先验**。许多高分辨率基准数据集 [43, 63, 86] 依赖多项选择格式，使模型能够利用纯文本先验来提高分数。**(4) 人工验证不足**。大量问答对由自动方式生成，缺乏严格的人工验证，导致题目脱离现实（例如，“图像中有多少辆汽车？”答案却是“300”）。

为弥补这些不足，我们提出了遥感高分辨率基准_（***RSHR-Bench***）——一个用于理解和复杂推理任务的超高分辨率遥感基准。RSHR-Bench整理了一个包含5,329幅全场景遥感图像的语料库，同时保留其原生分辨率：每幅图像的长边为 ≥4K，像素数最高可达 $\sim 3\times10^8$ （≈300百万像素）。图像来源涵盖广泛使用的超大场景数据集——DOTA v1.0/v2.0 [15, 71], XLRS-Bench [63], MiniFrance [9], FAIR1M [56], HRSCD [14]——以及我们采集的无人机拍摄影像（最高达 10^8 像素）。随后，我们设计了13个提示模板，并使用Qwen2.5-VL-7B [5]和GPT-5 Thinking [47] 生成四类任务：多项选择视觉问答、开放式视觉问答（自由形式）、图像描述和单图评估。对于视觉问答任务套件，我们设置了九个感知类别和四种推理类型，涵盖单图和多图场景。不同于以往侧重单轮选择题的遥感基准数据集，*RSHR-Bench* 还支持多轮对话和多图像对话，更贴近真实的遥感分析工作流程。

为验证数据质量，我们构建了一个半监督的人类—大语言模型验证流水线，以确保问题和答案正确、基于视觉，且不存在歧义或提示。第一阶段，我们采用纯大语言模型对抗式过滤，剔除无需查看图像即可解答的条目。随后，我们进行全面的人工审核，反复修改和复核，直至语言先验不再足以作答。为评估模型在原生分辨率下的整体理解能力，我们设置了图像描述任务，使用分辨率从 4K 到 10^8 像素的图像，并为每张图像提供十个视觉问答条目，以联合评估视觉问答和图像描述任务。随后，我们在 RSHR-Bench 中评估了多种模型，包括遥感视觉语言模型、通用视觉语言模型（开源和闭源）以及纯文本大语言模型。评估结果表明，所有十四种模型在四类任务上的表现都不佳。总体而言，我们的贡献可概括如下：

- 我们推出*RSHR-Bench*，这是一个全新的遥感基准，旨在公平、全面地评估视觉语言模型在超高分辨率图像上的表现。

- 我们设计了四种具有代表性的任务类型——选择题和开放式视觉问答、图像描述以及单图评估，用于评估通用视觉语言模型和遥感视觉语言模型的表现。

- 我们在 RSHR-Bench 上评估模型，发现所有模型的表现都不尽如人意，这凸显了推动模型进步、使其适用于现实世界遥感应用的必要性。

2. 相关工作

通用多模态基准。 近期的多模态基准推动了对视觉语言模型的量化评估，但其中许多基准聚焦于狭窄领域或少量任务（例如图像描述 [10, 50] 或视觉问答 [18, 20, 25, 44,54, 91]）。为扩大覆盖度，MME [3] 涵盖 14 项感知与认知任务，MMBench [37] 提供了涵盖 20 个技能维度的 3,000+个问题（例如定位和社会推理），Seed-Bench [72]进一步增加了问题数量（19k），而 MMT-Bench [77] 则引入了来自自动驾驶和嵌入式人工智能领域、面向应用的数据。与这些基准互为补充的是，若干高分辨率自然场景基准通过大尺寸输入对长上下文视觉进行压力测试，但仍不属于遥感领域：HRBench [68] 针对 4K/8K 级输入，TreeBench [66]评估 ∼2K 级输入下可追溯的视觉推理，而 MME-RealWorld [86]报告的典型图像大小约为 2,000×1,500，最大尺寸则非常高。然而，这些

通用数据集几乎不包含遥感图像，因此提供的遥感专属语义和标注有限。而且，其平均图像大小仍小于真实遥感场景（例如，HRSCD [17] 使用的图像最高达 100 百万像素）。

遥感多模态基准。现有的多模态模型遥感基准数据集主要分为三类。（*i*）中等分辨率下的图像描述/视觉问答：早期工作，如 RSIEval [64]和 LHRS-Bench [46]，旨在通过数百条人工整理的条目开展图像描述和视觉问答，主要关注视觉上下文有限的感知任务。（*ii*）可靠性与细粒度语义：RSSA [48]侧重于虚假信息检测，而 FIT-RSRC [42]和 VLEO-BENCH [81] 则着重于物体关系和场景级理解（例如城市监测、灾害救援、计数、定位、变化检测），将关注点从通用问答拓展到可信度和结构化语义。（*iii*）大规模与高分辨率：为适应现实中的遥感需求——检测任务中的图像分辨率通常超过 4,000×4,000（DOTA [71]），分割任务中甚至可达 10,000×10,000（HRSCD [17]）——近期基准数据集不断提升分辨率并拓展任务范围：XLRS-Bench[63]的图像平均分辨率为 ∼8,500×8,500，包含涵盖视觉问答、图像描述和定位的 16 个子任务；VRSBench [31] 提供 29,614 张图像，并包含丰富的问答和目标指代，支持多样化评估。MDAS [23] 则引入了高光谱分辨率，用于分类和变化检测。总体而言，该领域正从低分辨率图像描述和视觉问答，逐步转向更能满足现实场景部署需求的超高分辨率基准数据集。

多模态大语言模型。多模态大语言模型（MLLMs）在感知和推理方面发展迅速 [22, 51], , 然而，大多数通用系统（例如，LLaMA [58], Gemini [57], GPT-4o[26], Qwen-VL [4], InternLM-XComposer [82], MiniCPM[75], LLaVA [34], MiniGPT-4 [92]）并未针对超高分辨率输入进行专门优化，通常仅支持 2K–4K 图像。为克服这一局限，近期研究采用三种主要策略将多模态大语言模型扩展至高分辨率：通过全局对齐进行图像切块（LLaVA-Next [35]）、令牌／图像块压缩（Monkey[33]; LLaVA-UHD [19]），以及结合可学习查询的双编码器或多尺度融合（Mini-Gemini [32]; Cambrian[59]; SliME [76]）。这些策略在保留全局语义的同时缩短了序列长度。在遥感领域，研究主要聚焦于用于地理空间理解的多模态大语言模型 [7, 21, 24, 42, 65, 74, 76, 80, 88]。GeoChat [27] 针对遥感图像上的多轮对话进行指令微调；LHRS-Bot [46] 结合课程学习和志愿地理信息，引入多层次视觉—语言对齐，以增强依据能力；EarthGPT 则通过统一的生成式接口实现对多传感器（光学／合成孔径雷达／多光谱）数据的解译 [84]。在可靠性和规模方面，VHM [49] 减少幻觉并促进校准不确定性，而 GeoLLaVA-8K [62] 通过分块和上下文聚合实现超高分辨率（8K）推理。仍待解决的问题包括忠实的地理空间定位、多时相推理、大规模传感器融合以及不确定性调整。

3. 方法

整体流水线（图 2）包含三个阶段：数据集收集、问题生成，以及人工与大语言模型验证。我们还在第 3.1节讨论面临的困难，并与先前基准进行比较。在第 3.2节中，我们介绍四项主要任务的评估维度。

3.1. RSHR-Bench

数据集收集。我们将六个公开遥感基准数据集与一组高空无人机 ∼ 100百万像素 单次拍摄帧相结合，构建了一个高分辨率语料库。对于每个来源，我们均保留原生分辨率，并筛选长边至少为 4K 像素的图像。从 DOTA v1。0 [71] 和 DOTA v2。0 [15], 中，我们选取包含密集人造结构和远距离上下文的大幅面卫星场景。从 XLRS-Bench [63], 中，我们抽取单边达数万像素的超大全景图。从 HRSCD [14], 中，我们保留同一区域的配对视图，以支持时间一致性分析。从 MiniFrance [9], 中，我们选取纹理丰富、包含大量小物体的超高分辨率都市区域图块。从 FAIR1M 1。0 [56], 中，我们收集覆盖范围广的机场、港口和工业场景，其中包含密集的精细尺度目标，且观测角度多样。我们的无人机数据集会从不同视点和飞行姿态反复拍摄同一地理区域，每帧均以单次曝光采集，分辨率约为 100MP，从而保留原生几何结构。与典型的超高分辨率基准相比，后者的目标通常占据较大的像素范围 [8], , 我们的无人机场景则以小物体为主，并呈现显著的尺度和透视变化。总体而言，我们的图像收集遵循高分辨率、全面的场景覆盖、丰富的目标多样性以及面向任务等原则。包括无人机数据在内的汇总统计数据和分辨率分布见图 2。附录中提供了各数据集在不同分辨率下的数据收集流程和数量的详细说明。

问题生成。遵循近期的遥感基准数据集 [31, 63], , 我们设计了 13 个提示词模板，涵盖四种任务类型中的感知与推理，包括封闭式视觉问答（多项选择）、开放式视觉问答（自由形式）、图像描述和单图像理解。感知任务涵盖九类常见查询类别（例如颜色、方向/位置、区域定位、区域计数）。推理问题则针对特定图像，要求提供明确的视觉证据，并避免使用暗示正确选项的措辞（例如未来预测、目标状态判断）。设计中的所有任务均遵循——

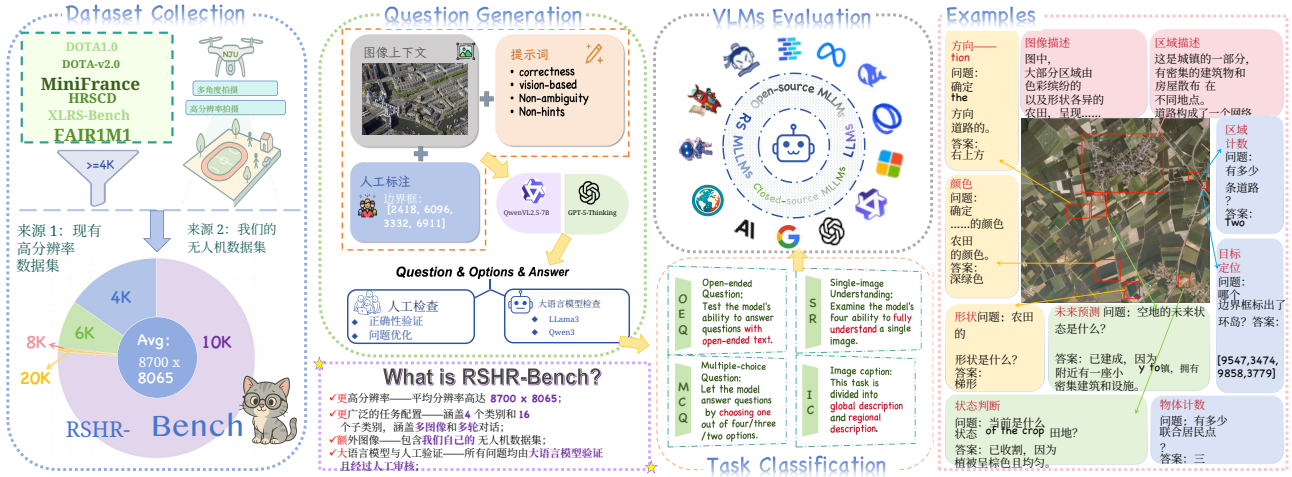


图2. 该概览展示了我们 RSHR-Bench 的建设过程：我们从多个数据集中收集高分辨率图像，并补充了来自自有无人机数据集的图像。随后，我们生成问题，并由大语言模型和人工进行验证。最终的任务分为四种主要类型，涵盖多种视觉语言模型评估实验。最后，右侧展示了单图像理解任务的示例，说明了该基准的应用方式。

以下三项原则——正确性、措辞明确无歧义和可通过视觉作答。对于封闭式和开放式视觉问答任务，我们优先选择数量少但具有显著特征的目标。这种方法着重考查高分辨率理解，同时确保标注过程中每个目标在上下文中都能被唯一识别。随后，我们通过两阶段起草流程设计这些问题及选项：对于提供了边界框和标签的感知类条目，使用 Qwen2.5-VL-7B 起草 [5], ；对于更难的组合类或多图像条目，则使用 GPT-5 Thinking 起草 [47]。对于全图图像描述，我们将每幅图像划分为四个方向扇区（上、下、左、右），并使用 GPT-5 Thinking [47] 生成分别描述各扇区的图像描述。对于单图像任务，我们为每幅图像设计十个问题，涵盖感知与推理，包括八个开放式问题、一条局部图像描述和一条全局图像描述。最终，我们生成了 50 幅图像和 500 个问题，用于评估模型的单图像理解能力。四类任务的所有问题和答案均由人工标注员撰写，而图像描述由 GPT-4o 生成，并由人工核验其正确性。所有提示词、模板和示例均见附录。

人工与大语言模型验证。 为确保问题质量，我们采用两阶段的人工—大语言模型验证流程。第一阶段将问题撰写与验证分开，以提高可靠性。标注员**A** 负责起草问题（并在适用时审查模型生成的问题与答案），确保满足四项标准：正确性、严格视觉基础对应、措辞清晰无歧义，且不包含无意间的提示。审阅员**B** 随后独立审核 A 的输出及任何模型草稿，标记不一致之处，并剔除无效条目。由六名受过培训的标注员（均拥有学士学位）组成的团队约用 300 小时完成了这一阶段。我们还保留了完整的溯源记录（包括图像标识符、提示词、模型输出和编辑历史），以支持追溯并消除歧义。第二阶段，我们使用纯文本大语言模型（Qwen3-8B [73] 和 Llama3-8B [45]）检测并排除仅凭语言先验即可解答的问题。具体而言，大语言模型在无法查看对应图像的情况下尝试回答每个问题。如果模型仅凭语言线索或统计规律答题时的准确率，高于依赖答案选项中的图像答题时的准确率，我们就会删除或修改受影响的条目，以减轻这一问题。我们反复改写，直至正确作答确实需要视觉定位，且所有选项都不含暗示性线索。经过多轮编辑和审查，我们最终获得了一个包含 1,932 组成对图像与问题的高质量数据集。完整的生成与验证流程耗费约 100 GPU小时。

流水线中的困难。 在我们的多任务设置中，构建可靠的监督和评估体系面临四项主要挑战。❶ **标注多样性。** 许多图像包含大量高度相似的目标，从而降低实例级多样性并助长过拟合。我们对语料库进行人工筛选，优先纳入更广泛的常见遥感场景和物体（例如土地利用类型、多种交通方式）。我们的目标是提高类间和类内差异。同时，我们致力于保留原生场景上下文，并在各数据划分中维持源分布。❷ **问题提示。** 描述性线索可能泄露标签（例如，森林 → 绿色），使模型能够依靠常识而非视觉证据进行排除。我们首先使用纯文本大语言模型评估提示词中的

语言依赖性。如果问题过度依赖语言，我们会通过人工后编辑移除词汇捷径并统一措辞，同时保留预期的视觉证据和难度。❸ **干扰项设计**。由于静态图像和分辨率的限制，为推理任务生成干扰项尤其具有挑战性，因为答案的解读往往并不唯一。例如，一片黄色的土地可能是自然要素，也可能是人工要素。然而，如果我们观察到作物倒伏的迹象，就可以推断这是自然灾害。为此，我们优先选择具有明确因果特征的物体，确保答案唯一且定义清晰。此外，我们将物体在现实世界中的独特特征融入答案设计，以减少歧义并提高可靠性。❹ **答案错误**。模型在小型边界框目标和计数任务上表现出系统性错误：即使正确答案是"0"，且选项中出现"0"，预测仍会偏向肯定；对于超高分辨率图像（$> 10^8$ 像素），检测召回率会急剧下降。我们通过有针对性的人工审核和修正来缓解这些影响，从而提高标签的准确性和评估公平性。

与先前基准的比较。如表1所示，我们将本基准与现有遥感数据集进行了比较。具体而言，我们考虑两类遥感数据集：非视觉问答和视觉问答。非视觉问答数据集（例如，DOTA [15, 71], DIOR [29]）侧重于识别或图像描述，通常缺乏多轮对话和多图像推理。现有视觉问答数据集主要为低分辨率数据集（通常为 512×512），并且局限于单轮、单图像交互（例如，RSVQA [38], HRVQA [28], RSIVQA [90]）。相比之下，我们的基准强调：❶ 更高分辨率的图像。我们整理了最高达10^9 像素的超高分辨率图像，将评估推向真实世界的遥感场景。我们的基准将超高分辨率图像与交互丰富的评估和明确的可靠性检查相结合：图像覆盖数千万至数亿像素（平均值为 8,700 × 8,065；最大值为 29,200 × 27,620），来源包括卫星、航空和无人机。❷ 大语言模型验证。在刻意降低分辨率的视图下，大语言模型对感知和推理题目进行验证，以强调视觉能力而非语言先验。我们采用人工–大语言模型验证流水线，并结合基于大语言模型的验证，构建了一个经过筛选且具有挑战性的视觉问答集（涵盖 1,366 张图像中的 1,578 个题目），针对此前仅得到部分覆盖的感知到推理链。❸ 全局图像感知和更广泛的任务配置。我们改变交互形式（例如，单图多轮和多图单轮），以探究多图像融合、记忆和决策，并将任务覆盖范围从细粒度感知（颜色/形状线索）扩展到跨图像推理（多区域对比、物体状态判断和未来预测）。

表1. 现有遥感基准与我们的基准之间的比较。✗和✓分别表示未包含和已包含，◯，✓和◯✓ 分别表示机器生成、人工撰写和半自动（即先由机器生成，再经人工验证）。上：非视觉问答；下：视觉问答。短横线（–）表示缺失或未报告。

数据集	来源	图像	平均分辨率	最大分辨率	多轮标注方式		
非视觉问答							
GeoPixel [52]	航空	65,463	2,560×2,560	4,960×4,960	✗	✓	✗
refGeo [83]	航空、无人机	80,000	640×640	1,024×1,024	✗	✓	✗
LuoJiaHOG [89]	TBD	94,856	640×640	1,024×1,024	✗	✓	✗
DOTA测0 [15]版	航空、卫星	2,806	4,000×4,000	20,000×20,000	✗	✓	✗
FAIR1M1.0 [56]	航空、卫星	11,268	4,500×4,500	20,000×20,000	✗	✓	✗
RSICD [40]	卫星	15,000	5,300×5,300	10,000×10,000	✗	✓	✗
RSVG [70]	航空,卫星	10,921	224×224	224×224	✗	✓	✗
RRSIS-D [78]	卫星	17,402	800×800	800×800	✗	✓	✗
RSIEval [64]	航空,卫星	17,402	800×800	800×800	✗	✓	✗
DIOR [29]	卫星	100	512×512	512×512	✗	✓	✗
MillionAID [30]	航空，卫星	23,463	1,200×1,200	2,000×2,000	✗	✓	✗
RSICap [41]	航空，卫星	1,000,848	—	31k×31k	✗	✓	✗
		2,585	224×224	224×224	✗	✓	✗
VQA							
XLRS-Bench [63]	卫星，航空	3,079	8,500×8,500	10,000×10,000	✗	✓	✗
LRS-VQA [43]	卫星,航空	1,657	5,403×4,935	26,176×24,832	✗	✓	✗
RSIVQA [90]	航空,卫星	37,264	512×512	512×512	✗	✓	✗
VRSBench [31]	航空,卫星	29,614	512×512	512×512	✓	✓	✗
RSVQA [38]	航空,卫星	11,431	499×499	512×512	✗	◯	✗
HRVQA [28]	航空	53,512	1,024×1,024	1,024×1,024	✗	◯	✗
TAMMI [85]	航空，卫星	282,852	—	—	✗	◯	✗
RSVLM-QA [93]	航空，卫星	6,000	512×512	512×512	✗	✓	✗
EarthVQA [67]	卫星	6,000	512×512	512×512	✗	✓	✗
SegmentationVQA [60]	航空	16,274	2,560×2,560	4,096×4,096	✗	◯	✗
RSHR-Bench	卫星、航空、无人机	5,329	8,700×8,065	29,200×27,620	✓	✓	✓

3.2. 评估维度

概览。我们沿着两个互补维度评估模型——感知 和推理（图3 （左））——通过四项任务（图 3 （右））评估模型在超高分辨率遥感图像中的识别、定位、依据和整体理解能力。（i）多项选择视觉问答：评估模型在固定答案空间内的决策能力。（ii）开放式视觉问答：该任务不提供选择题选项，用于评估自由形式的视觉理解和组合性，从而更准确地衡量模型的能力。（iii）图像描述：要求对区域和完整图像进行简洁、准确的描述。（iv）单图评估：评估单图理解能力，涵盖不同分辨率下的感知和推理，以及图像描述。

感知。我们的基准将感知分解为一系列局部和场景级任务，用于检验高分辨率遥感图像中的识别、定位和空间推理能力。具体而言，我们评估：(i) 颜色检测，要求识别并分类给定边界框内的颜色；(ii) 形状/边缘识别，考察模型精确描绘目标边界的能力；(iii) 方向检测，评估模型根据裁剪区域估计目标朝向或延伸方向的能力；(iv) 物体分类，涵盖土地利用类别以及航空场景中常见的细粒度目标（例如码头、泊位区和船舶类型）；(v)目标空间关系，引导模型对多个实体之间的相对位置进行结构化描述；(vi) 目标定位，衡量模型根据精确的自然语言指称进行定位的准确率；(vii) 区域定位，聚焦于组合式子区域（例如三条道路与附近鱼塘交汇的区域）；以及 (viii)计数，包括

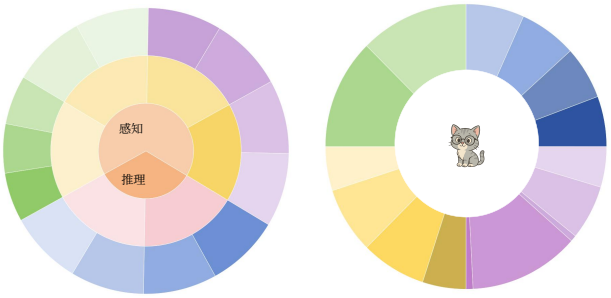


图3. 基准测试构成概览。左侧：感知与推理的任务类别。右侧：四个能力组中的任务数量——多项选择视觉问答（MCQ）、开放式问题（OEQ）、图像描述（IC）和单图评估（SIE）。

全局图像层面的计数，以及在给定边界框内进行区域计数。

推理。 作为对感知能力的补充，我们的推理赛道旨在评估时间、因果和上下文推断能力。我们纳入了 (i)异常检测与解释，要求模型不仅定位异常或显著区域，还要推断可能的原因（例如清理、洪水、建设、森林砍伐），并应对单图和多轮设置；(ii) 未来预测，通过两种配置进行评估：单图多轮（50 组，每组 3 个问题）和多图单轮（50 对跨年份、地点相同的图像），要求根据当前证据推演可信的发展变化；(iii)多区域联合对比，比较并综合一张或多张图像中多个区域的信息；以及 (iv) 目标状态判断，根据上下文线索判断目标是静态还是动态。通过在一致的输入和提示词下统一这些任务，该基准全面衡量模型将“什么在哪里”与“这意味着什么以及接下来会发生什么”联系起来的能力，为力求实现真实世界遥感智能的视觉语言系统树立了更高标准。

视觉问答任务套件。我们通过九项感知和四项推理任务评估视觉语言能力，这些任务定义在单图像和多图像设置中，并涵盖单轮和多轮交互。遵循近期基准数据集 [43, 63], , 我们采用多项选择视觉问答（A/B/C/D）评估所有视觉语言模型。为减轻多项选择格式固有的语言先验和随机猜测问题，我们进一步采用标准化流程进行开放式转换：对于每个条目，GPT-4o [26] （i）移除选项；（ii）改写问题，使答案必须依据视觉证据而非文本线索或常识先验推断得出；（iii）确保不检查图像就无法回答改写后的问题。每个改写后的条目都经过人工审核，以确认视觉依据明确且表述清晰。总计，我们为 1,366 张图像中的 1,578 个物体标注了边界框，并生成了 1,932 个问答对。因此，开放式视觉问答设置不提供答案线索，因此比其多项选择对应形式更具挑战性。

图像描述任务。 对于遥感数据集， [38, 43, 63], 图像描述是一项标准评估任务，可检验模型的场景理解能力。我们汇集了一个超高分辨率图像语料库（分辨率为 $\geq$4K；场景主要为 $\sim$100 百万像素），涵盖港口、机场、森林以及城市和农村区域。我们首先设计提示模板，并在单张图像上反复验证，以确保图像描述生成的稳健性。随后，我们使用 *GPT-5 Thinking* 生成：(i) 全局摘要，概述主要场景、显著物体、空间布局和功能用途；以及 (ii) 结构化的区域描述，分别对应上、下、左、右区域——每段描述均以区域级概述开头，随后提供细粒度细节，并明确提示模型描述属性（颜色、位置、大小、形状）和空间关系。生成指南要求避免使用模糊的量词（例如“很多”“一些”“少数”）。我们总共为 3,913 张图像生成了参考描述，并随机抽取 100 张进行人工核验，以评估正确性并减少幻觉。我们使用参考评判器评分，*GPT-5Thinking* [47], 并通过迭代优化提示词和指令，使评判分数更好地契合人工评估。评估标准强调：(i) 视觉保真度和依据；(ii) 属性与关系的覆盖度；(iii) 连贯流畅且无冗余；以及 (iv) 事实一致性高、幻觉率低。此外，我们使用 GPT-4o [26]评估其他模型生成的描述与 GPT-5 Thinking 生成的描述之间的分数 [47],, 并分别评估全局（整幅图像）摘要和区域（上/下/左/右）描述。

单图评估。 我们设计了一套单图评估协议，用于测试模型是否真正理解单张遥感图像。在此设置下，我们选取了 50 张分辨率从 $\geq$4K 到 $\sim 2\times10^8$像素不等的图像，并为每张图像标注了 10 个代表性子任务，涵盖四个能力组。所有问答对均由至少拥有学士学位的人工标注员创建并仔细核验。这些子任务涵盖区域级和图像级理解（图像描述、区域描述），以及对基本物体属性的感知（颜色、形状/边界、方向、物体分类）。此外，还包括对物体关系的推理，以及具有挑战性的定位或计数（空间关系、物体和区域定位、物体和区域计数）；也涵盖基于视觉证据和遥感知识的高阶能力（状态判断、异常检测、未来预测）。

4. 实验

指标。 对于多项选择视觉问答，我们通过比较模型预测的选项与标准答案（

表2. 视觉感知、推理和多轮任务的结果。从左到右依次为：感知——COL=颜色检测、SHP=形状识别、DET=检测、OC=物体分类、REL=物体空间关系、OGD=物体依据、RG=区域定位、OCN=物体计数、RCN=区域计数、Avg.=感知平均值；推理——AD=异常（单轮）、FP=未来预测（多图像）、MRJC=多区域联合对比（多图像）、MRJCS=多区域联合对比（单图像、多边界框）、OSJ=物体状态判断（单轮）、Avg.=推理平均值；多轮——MAD=异常、MTFP=未来预测、MOSJ=物体状态判断；高亮规则：按列计算的最大值（不包括两行大语言模型）以蓝色显示；按行计算的最大值——仅在感知和推理任务列中计算（不包括两个平均值列、所有多轮/MTEM@1，以及两行大语言模型）——以红色显示。并列值均高亮；若同时符合两条规则，则优先显示为蓝色。https://github.com/Yunkaidang/RSHR

模型	感知									Avg.	推理					Avg.	多轮			MTEM@1
	COL	SHP	DET	OC	REL	OGD	RG	OCN	RCN		AD	FP	MRJC	MRJCS	OSJ		MAD	MTFP	MOSJ	
遥感视觉语言模型																				
EarthDial [55]	41.0	22.0	21.0	30.0	32.5	30.5	27.1	18.0	31.0	28.1	42.0	30.0	29.5	32.0	52.0	37.1	56.7	60.0	73.5	30.7
GeoChat [27]	32.5	22.0	24.0	29.5	40.0	25.0	22.9	22.5	29.0	25.9	30.0	24.0	25.5	30.0	32.0	28.3	48.3	46.0	62.9	19.3
GeoLLaVA-8K [62]	25.0	24.0	25.0	25.0	25.0	21.4	25.0	25.0	24.5	24.0	0.0	0.0	34.0	22.0	16.0	25.0	24.7	47.7	7.0	
VHM [49]	25.5	25.0	26.0	26.5	55.0	25.0	22.9	25.0	25.0	25.7	26.0	24.0	26.5	34.0	28.0	27.7	45.0	53.3	46.2	16.7
开源视觉语言模型																				
InternVL2.5-8B [11]	25.5	22.0	26.0	26.0	22.5	24.5	30.0	22.5	20.0	24.3	26.0	20.0	22.5	34.0	20.0	24.5	25.0	28.7	35.6	1.8
InternVL 3.5 8B [69]	21.5	28.0	18.0	21.5	29.0	28.5	30.0	26.5	25.0	25.3	20.0	16.0	29.0	34.0	26.0	25.0	30.0	22.7	40.2	6.1
MiniCPM2_6 [75]	21.5	28.0	30.0	24.0	19.5	29.5	34.3	22.0	29.0	27.4	26.0	30.0	35.0	32.0	30.0	30.6	26.7	23.3	31.1	1.8
Phi-3.5-视觉 [1]	25.0	24.0	25.0	25.0	23.5	25.0	22.9	25.0	25.0	24.5	24.0	22.0	23.5	30.0	22.0	24.3	28.3	24.7	47.0	7.0
Qwen2.5-VL-7B [5]	29.5	25.0	22.0	28.0	25.0	24.5	24.3	26.5	22.0	25.2	26.0	28.0	25.0	10.0	20.0	21.8	21.7	24.0	10.6	0.0
DeepSeek-VL [39]	22.5	22.0	21.0	25.0	20.5	26.0	28.6	20.5	22.0	23.1	22.0	28.0	50.0	32.0	20.0	30.4	20.0	23.3	33.3	9.1
VILA-HD [53]	40.0	22.0	22.0	37.0	35.5	26.0	21.4	24.5	24.0	28.0	58.0	30.0	55.0	32.0	58.0	46.6	65.0	57.3	57.6	24.8
闭源视觉语言模型																				
GPT5 [47]	29.0	10.0	23.0	23.0	37.0	24.5	31.4	20.0	23.0	24.5	74.0	58.0	35.0	34.0	66.0	53.4	78.3	73.3	86.4	52.6
GPT-4o [26]	49.5	23.0	15.0	35.5	30.5	28.0	27.1	22.5	41.0	30.2	68.0	56.0	30.5	32.0	64.0	50.1	70.0	72.0	84.1	47.4
GPT-4o-mini [26]	41.5	16.0	29.0	31.5	31.5	32.0	28.6	19.5	32.0	29.1	54.0	54.0	31.5	48.0	54.0	48.3	78.3	68.0	75.0	41.2
Gemini-2.5-pro [12]	55.0	18.0	31.0	40.0	41.5	32.5	45.7	25.0	25.0	34.9	66.0	32.0	41.5	38.0	50.0	45.5	56.7	60.0	57.6	32.6
LLMs																				
Llama3-8B [45]	23.0	22.0	35.0	22.0	27.5	25.0	22.9	21.5	29.0	25.3	30.0	30.0	27.5	34.0	48.0	33.9	58.3	50.0	60.6	16.7
Qwen3-8B [73]	38.5	28.0	26.0	36.5	31.0	24.5	30.0	25.5	48.0	32.0	42.0	26.0	31.0	56.0	56.0	42.2	53.3	58.0	57.6	17.5

A/B/C/D）来报告准确率。对于开放式视觉问答，我们使用 GPT- 4o [26] 作为自动评审，对每个回答按 1–100 分打分，衡量其与人工标注参考答案的一致程度。评分遵循专家评审标准：优先考虑与参考答案的一致性，接受语义等同的表述以及适度的数值／单位误差，仅对实质性幻觉予以扣分。得分达到 ≥ 80 的回答计为正确。遵循近期研究 [31, 63], ，我们报告 BLEU-1、2、3 和 4、METEOR 及 ROUGE-L，以评估图像描述任务的描述质量。对于多轮评估，我们报告对话级完全匹配率（MT-EM%）。令 $\mathcal{D}$ 为评估对话的集合。对于每段对话 d，令 n_d 为有效轮次的数量；对于每个轮次 $i = 1, \ldots, n_d$，定义每轮正确性$z_{d,i} \in [0, 1]$ 为 $z_{d,i} = \mathbf{1}[\hat{y}_{d,i} = y_{d,i}]$（用于离散选择），或 $z_{d,i} = s_{d,i}/100$（用于原始分数为$s_{d,i} \in \{1, \ldots, 100\}$的评分评估）。统一 MTEM@1 和 MTEM@80 的严格对话级“全部正确”指标为

$$\mathrm{MTEM@}t = \frac{100}{|\mathcal{D}|} \sum_{d \in \mathcal{D}} \mathbf{1}\left[\min_{1 \leq i \leq n_d} z_{d,i} \geq t\right]. \qquad (1)$$

此处 $t \in \{1, 0.8\}$ 分别对应 MTEM@1 和 MTEM@80（即，对于 t=1，所有有效轮次都必须完全正确；或者对于 t=0.8，每个轮次在原始 1–100 分制下的得分都必须至少为 80 分）。

RSHR-Bench（选择题）上的结果。如表2所示，我们评估了四个模型系列（包括遥感视觉语言模型、开源视觉语言模型、闭源视觉语言模型和纯文本大语言模型），涵盖九项感知子任务、五项推理子任务和三项多轮任务。我们发现，Gemini-2.5-pro [12] 在低级感知方面准确率最高，而 GPT5 [47] 的整体表现最佳（AD/FP/OSJ = 74.0/58.0/66.0）。开源模型的准确率大多徘徊在 25% 左右，并且难以应对组合推理任务。GeoLLaVA-8K [62]只能处理选择题，且严重偏好选项 A，因此 MRJC 和 MRJCS 的准确率均为 0。值得注意的是，支持 4K输入的 VILA-HD [53], 在推理任务上的表现显著优于其他开源模型（平均值 58.0）。在多轮对话方面，GPT-5 表现最强（MAD/MTFP/MOSJ = 78.3/73.3/86.4；MTEM@1 = 52.6）。即使经过多轮人工审核，确保所有问题都需要视觉证据，纯文本大语言模型（Qwen3-8B [73] 和 Llama3-8B [45]）在推理任务上的准确率仍超过 30%。

RSHR-Bench（开放式）上的结果。为避免模型利用多项选择格式中的选项先验，我们将每个条目转换为不含选项的开放式问题。表3 报告了各任务准确率，以及在 80/100 阈值下的多轮精确匹配率（**MTEM@80**）。如图4（左）所示，所有模型在感知任务上的得分均低于 50，表明其几乎无法给出正确且相关的回答，也凸显了模型在超高分辨率图像上的感知能力不足。在推理任务上，闭源模型和开源视觉语言模型的平均得分均在 50 左右，但在

表 3. **视觉**感知、推理**和**多轮**任务的结果**。 任务缩写见表 2。 **平均值 (**感知/推理)分别表示所有感知/推理任务列的平均值。 高亮说明：按列计算的最大值 (不包括两行大语言模型) 以蓝色标示；按行计算的最大值 —— 仅在**感知**和**推理**任务列中计算 (不包括两个平均值列、所有多轮/MTEM@80，以及两行大语言模型)以红色标示。并列最大值也会高亮；若同时符合两条规则，则蓝色优先。

模型	感知									Avg.	推理					Avg.	多轮			MTEM @80
	COL	SHP	DET	OC	REL	OGD	RG	OCN	RCN		AD	FP	MRJC	MRJCS	OSJ		MAD	MTFP	MOSJ	
遥感视觉语言模型																				
EarthDial [55]	34.3	35.0	42.4	50.8	52.7	28.6	25.5	24.0	45.3	37.6	55.4	62.6	58.0	58.7	57.9	48.7	65.2	65.7	65.9	11.4
GeoChat [27]	35.7	40.5	41.1	47.0	54.9	24.6	22.6	35.5	44.2	38.5	49.8	39.9	24.5	24.0	37.6	41.5	53.9	58.3	55.3	2.6
VHM [49]	49.3	42.9	39.6	63.0	48.2	24.4	23.5	29.1	54.3	41.6	60.6	62.7	38.8	35.2	51.4	48.7	60.9	63.1	71.3	16.7
开源视觉语言模型																				
InternVL2.5-8B [11]	53.2	37.6	31.6	63.6	59.1	26.9	30.1	46.1	41.8	43.3	74.3	74.7	58.3	48.5	61.8	52.6	67.6	71.3	72.4	24.8
InternVL3.5-8B [69]	37.1	44.6	36.8	68.1	54.2	26.9	31.7	41.1	34.3	41.6	66.0	67.6	68.0	43.0	58.6	51.5	70.0	71.1	66.7	19.3
MiniCPM2_6 [75]	49.3	38.6	37.8	60.8	58.9	23.2	29.0	35.0	33.0	40.6	71.3	70.9	51.8	43.1	60.2	51.7	63.5	75.1	74.1	30.7
Phi-3.5-视觉 [1]	40.4	32.2	39.8	61.1	56.5	28.9	26.6	36.3	32.3	39.3	67.5	67.1	43.6	40.1	55.2	50.3	64.5	70.8	75.1	22.8
Qwen2.5-VL-7B [5]	35.1	34.6	30.0	55.5	57.9	26.0	30.2	33.2	28.3	36.8	58.8	37.4	67.1	33.5	51.4	40.2	55.9	49.6	54.2	2.6
DeepSeek-VL [39]	54.1	35.2	34.2	63.5	60.2	25.4	24.5	38.5	33.9	41.1	70.4	60.1	46.8	35.2	55.1	49.5	67.7	68.4	60.2	19.3
VILA-HD [53]	58.4	40.7	33.4	56.8	67.8	27.7	29.6	43.9	37.9	44.0	62.3	60.1	62.4	32.5	56.3	53.0	71.1	74.1	78.6	67.5
闭源视觉语言模型																				
GPT5 [47]	25.6	26.6	37.6	49.0	52.4	20.9	29.9	34.8	27.4	33.8	70.7	62.5	48.2	44.6	60.1	48.0	75.0	76.8	59.8	72.8
GPT-4o [26]	55.6	35.0	33.8	56.5	46.6	29.9	35.4	35.1	40.8	41.0	74.4	41.0	28.6	55.8	53.8	49.6	54.5	74.9	76.0	68.9
GPT-4o-mini [26]	29.3	37.6	31.8	62.2	45.9	28.4	23.0	30.1	24.3	34.7	70.5	72.2	65.0	43.9	63.9	51.0	78.0	74.9	82.3	72.5
Gemini-2.5-pro [12]	34.4	25.5	23.3	38.0	51.2	18.4	17.3	25.4	16.6	27.8	42.7	30.8	60.4	29.0	41.7	29.6	35.0	20.3	41.9	20.6
LLMs																				
Llama3-8B [45]	21.2	31.0	28.7	28.6	20.2	28.7	28.6	20.3	26.3	26.0	42.5	33.6	18.2	21.6	28.6	28.6	34.8	38.5	35.3	0.0
Qwen3-8B [73]	26.3	29.1	25.7	54.8	34.1	23.2	24.1	18.2	21.9	28.6	65.9	68.1	21.6	22.6	45.0	38.5	57.8	57.2	57.3	5.3

表4. 文本生成指标的实验结果。闭源模型以灰色标出。

方法	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE-L
InternVL2.5-8B [11]	11.0	4.7	1.7	0.6	17.0	17.3
InternVL 3.5 8B [69]	33.7	14.4	6.1	2.5	26.4	23.5
MiniCPM2_6 [75]	2.9	1.2	0.5	0.2	11.8	13.6
Phi-3.5-视觉 [1]	36.5	14.0	5.6	2.1	25.1	22.3
Qwen2.5-VL-7B [5]	0.1	0.0	0.0	0.0	3.8	6.1
DeepSeek-VL [39]	2.7	1.2	0.5	0.2	11.8	14.6
VILA-HD [53]	28.0	10.0	4.2	1.7	21.6	21.8
GPT5 [47]	43.3	15.3	5.0	1.6	30.1	20.6
GPT-4o-mini [26]	49.7	21.8	10.0	4.8	33.7	25.8
GPT-4o [26]	43.8	18.1	7.9	3.6	29.2	23.1
Gemini-2.5-pro [12]	9.1	3.0	0.9	0.3	7.5	7.1

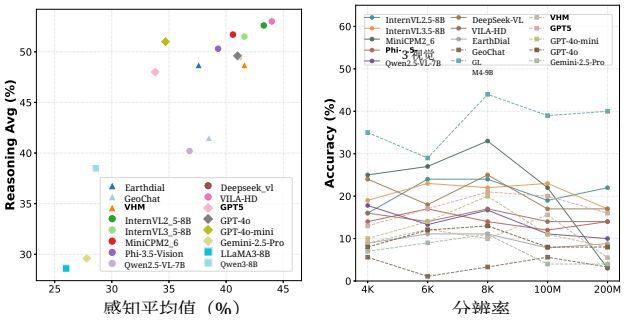


图4. 不同模型在 RSHR-Bench 上的实验结果。左：模型在开放式视觉问答中的表现。右：模型在单图评估中的表现。

MRJC（多区域联合对比，多图像）和 MRJCS（多区域联合对比，单图像、多边界框）上的表现明显下降。在遥感视觉语言模型中，开源和闭源视觉语言模型得分相近。对于多轮对话，闭源模型表现出色：GPT 5 [47], GPT-4o [26], 和 GPT-4o-mini [26], 以及开源 VILA-HD，在对话级指标 MTEM@80 上表现良好。Qwen3-8B [73]和 Llama3-8B [45], 的平均得分约为 30，这表明其回答不正确，并且存在大量幻觉。

图像描述任务的结果。如表6所示，我们评估了图像描述任务中广泛使用的指标。结果表明，在所有文本生成指标上，闭源视觉语言模型均优于开源基线。GPT-4o-mini [26] 取得了最佳综合成绩——BLEU- 4（4.8）、METEOR（33.7）和 ROUGE-L（25.8）。值得注意的是，GPT5 [47] 的 BLEU-1（43.3）和 METEOR（30.1）得分较高，但 BLEU-4（1.6）相对较低，这表明其描述可能采用了释义或句法更自由的表达，从而降低了精确的 n 元语法重叠。Gemini-2.5-pro [12] 和 Qwen2.5- VL-7B [5] 的表现差距较大（后者的 BLEU 接近于零），这表明在我们的评估设置下可能存在领域不匹配。

单张图像评估结果。如图4 （右）所示，我们在实用的高分辨率（4K–8K）和超高像素区间（100M/200M）下评估了所有模型。性能对分辨率高度敏感：在 4K–8K 下准确率一般，在 100M/200M 下则显著下降。无论是开源和闭源通用模型，还是遥感视觉语言模型，结果都普遍偏低（约 30%），表明它们难以稳健应对极高像素数量和广阔的空间上下文。总体而言，当前视觉语言模型在可靠理解高分辨率遥感图像方面仍面临挑战。

5. 结论

我们推出了**RSHR-Bench**，这是一个面向超高分辨率遥感图像视觉-语言理解的大规模基准。RSHR-Bench 保留原生分辨

率，最高可达 $\sim 3 \times 10^8$ 像素，并对通用视觉语言模型和遥感视觉语言模型进行全面、公平的评估。针对广泛的开源和闭源模型开展的实验显示，它们的表现普遍较低。我们希望 RSHR-Bench 能成为一个具有挑战性的基准，推动未来模型弥合这一差距，服务于现实世界中的遥感应用。

6. 附录概览

本附录补充介绍所提出的**RSHR-Bench**，并提供因篇幅限制而未在正文中详述的内容。附录其余部分安排如下。

第 7 节介绍人工标注流程和评估方案，包括两阶段人工标注与审核流程，以及人工评估。

第8节补充介绍基准设计的更多细节，包括任务分类体系和问答格式、不同模型的输入分辨率策略、图像描述评估以及无人机数据源。

第 9 节报告了其他基准数据集上的更多结果，包括 MME-RealWorld、XLRS-Bench和LRS-VQA，并分析了对输入分辨率的鲁棒性。

最后，第 10 节展示了更多案例研究，以及各项任务的详细提示模板。

7. 人工标注与评估

标注阶段的详细信息。该流程分为两个阶段。第一阶段，三名本科生花费约 50 人时，为部分问题绘制边界框并撰写答案。对于单图设置，我们首先根据长边长度将候选图像分为五个分辨率区间：4K、6K、8K、100M 和 200M 像素（例如，4K 区间包含长边至少为 4,000 像素的图像，其余区间以此类推）。不同任务会按依赖任务而异的比例从这些区间中抽取图像：形状识别或边缘识别等本质上依赖精细空间细节的任务，会偏向较高分辨率的区间。在每张选定的图像中，标注人员会选择可用于多项任务且难度分布有梯度的目标物体，例如通过改变物体像素面积与整张图像像素面积的比例来体现难度。第二阶段，由同一项目中的另一组本科生（与前三名标注人员互不重叠）进行第二轮审核。他们会检查所有已标注条目，并修改或丢弃可能存在歧义或事实错误的条目。这一流程确保了正确性和严格视觉基础对应。经过这两个阶段后，每种已标注任务类型都包含约 100–200 个经过验证的条目，并且各分辨率区间的图像组成受到控制，覆盖五个区间。

人工评估详情。 我们对精心整理的问题集开展人工评估，以估算人类答题的准确率。具体而言，我们设计了 17 种任务类型，分为感知（颜色（ COL ）、形状（ SHP）、检测（DET）、物体分类（OC）、关系（REL）、目标定位（OGD）、区域定位（RG）、对象计数（ OCN）、区域计数（RCN））、推理 （异常检测（AD）、基于两幅图像的未来预测（FP）、多区域联合对比（ MRJC）、单轮多区域联合对比（MRJCS）、目标状态判断（OSJ））和多轮（多轮异常检测（MAD）、多轮未来预测（MTFP）、多轮目标状态判断（MOSJ））。每种任务抽取 15 个图像—查询实例，共计 255 个实例。向人工标注员展示具体任务描述、输入图像和问题，并要求他们在看不到参考答案或模型预测的情况下作答。对于每个问题，评分者均不同于参与该问题构建的原始标注员和审核者。各任务的准确率按判定为正确的实例比例计算（例如，COL/SHP/DET/REL/RG/OCN 均为 15/15，OGD 为 14/15，OC 和 RCN 为 13/15，AD 为 10/15，FP 为 12/15，MRJC 和 MRJCS 均为 15/15，OSJ 为 14/15，MAD 和 MTFP 均为 13/15，MOSJ 为 15/15）。总体而言，模型在 255 个实例中答对了 237 个，对应的人工评估平均准确率为 92.94%。

人工与大语言模型验证的详情。对于所有任务，我们首先通过以下方式获取候选问题和选项：向模型（ Qwen2.5-VL-8B 和 GPT-5 Thinking）提供人工标注的边界框和标签，或直接向这些模型输入图像和提示词，让它们拟写问答对。随后，我们采用两阶段验证流程。第一阶段，人工标注员会检查每个问题、选项集和答案，并纠正任何事实错误，或图像与标注答案不符之处。第二阶段，我们进行纯文本验证：要求大语言模型在无法查看图像的情况下回答每个问题。如果模型在这种设置下的准确率较高（通常超过 30%），我们便认为仅凭语言就能过于轻易地解答该题，并以模型解释为指导进行修改。修改时，我们会去除词汇提示，并减少对常识先验的依赖；例如，对于“确定边界框内飞机的颜色”这一提示词及其答案“白色”，我们会将“飞机”替换为“主体对象”等中性表述。我们还会纠正纯文本测试中观察到的标签偏差：例如，在方向任务中，Qwen3-8B 往往倾向于选择“右上方”，因此我们可能会将正确方向改为“左下方”，或改选其他对象。这些调整会降低纯文本准确率，通常使其低于 30% 的阈值，或降至实际上难以进一步降低的水平。总之，这一

两阶段流程既确保了严格的视觉依据，又维持了经过妥善校准的任务难度。

8. 我们基准测试的详情

输入分辨率。表5总结了受评模型的最大输入分辨率或官方公布的分辨率。只有 GeoLLaVA-8K 和 Claude3.7 明确支持最高达 8000×8000的超高分辨率输入。大多数开源视觉语言模型受限于分块处理，输入分辨率约为 448×448。对于闭源模型，提供商并未透露确切限制，因此我们将其分辨率标记为未披露。许多遥感模型也依赖基于 CLIP 和 ViT 的视觉编码器，实际处理的分块输入分辨率约为 336×336。不过，由于原始论文通常没有明确说明这一策略，我们谨慎地将这些情况标记为未披露。

表5. 各模型公布的最大图像尺寸。

模型	最大输入分辨率
GeoLLaVA-8K [62]	8000×8000
GeoChat [27]	未披露
EarthDial [55]	未披露
VHM [49]	未披露
InternVL3-8B [69]	448×448切片
QwenVL2.5-7B [5]	动态标记化
Claude3.7 [2]	8000×8000
GPT-4o [26]	未披露

图像描述。我们还在一个包含 3,913 张图像的数据集上评估了图像描述性能。出于效率考虑，论文正文报告了一个包含 1,000 张图像的子集上的结果，因为调用闭源模型的商业 API 成本较高。3,913 张图像数据集上的完整结果见补充材料。对于每张图像，我们首先使用 GPT-5-Thinking 生成参考描述，然后计算模型输出与这些参考描述之间的标准图像描述指标——BLEU（1–4）、METEOR 和 ROUGE-L。表6 总结了结果，分数越高表示描述质量越好。

表6. 图像描述指标的实验结果。

方法	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE-L
InternVL2.5-8B [69]	33.3	14.4	6.2	2.8	26.2	24.1
InternVL3.5-8B [69]	33.4	14.3	6.0	2.4	26.2	23.4
DeepSeek-VL [39]	2.7	1.2	0.5	0.2	11.8	14.6
GLM4-9B [16]	37.3	11.5	3.5	1.1	23.8	19.0
MiniCPM2_6 [75]	2.9	1.2	0.5	0.2	11.7	13.6
Phi-3.5-Vision [1]	36.5	14.0	5.6	2.1	25.1	22.3
Qwen2.5-VL-7B [5]	0.1	0.0	0.0	0.0	3.8	6.1
VILA-HD [53]	28.1	10.0	4.2	1.7	21.6	21.8

任务分类体系与问答格式。表7 总结了 RSHR-Bench 的分层任务设计，涵盖从高层 L1类别（感知、推理、图像描述）到更细粒度的 L2/L3任务。对于每项任务，我们列出了任务代码、标注流程（全人工或半自动化）、问题生成器、交互模式（例如单图/多图、单轮/多轮或多框定位）、答案格式（选择题和/或开放式）以及问题实例数量。该分类体系涵盖低层感知技能（例如颜色、形状、计数）、更高层次的逻辑与知识依据推理（例如异常解读、未来预测），以及整体图像描述和区域描述，从而全面评估遥感视觉语言模型在感知和推理方面的能力。

无人机数据源。我们收集了由无人机在户外公共空间拍摄的超高分辨率数据。所有数据均采集于一座公共公园及附近的无人机作业场地。每段序列均在一次连续飞行中拍摄完成，飞行高度介于 120 米至 150 米之间。无人机沿平滑轨迹飞行，并在较长时间内观测同一区域。摄像机从多个朝向拍摄场景，既有仰视视角，也有俯视视角，因此这些帧构成了同一场景的长时间连续多视角视频流。每帧约包含 10^8 像素（约 1 亿像素），场景主要为公园和无人机作业场地。我们使用这些图像执行图像描述和视觉问答任务，连续的多视角结构为时间推理和多视角推理提供了丰富的上下文。

9. 其他基准数据集上的结果

MME-RealWorld 上的结果。如表 8所示，我们在 MME-RealWorld 的遥感子集上评估了若干语言模型和未输入图像的多模态模型；该评估中，所有图像均被移除，模型仅接收文本输入。令人惊讶的是，表现最佳的纯语言模型（Llama3-8B）仍能正确回答 31.22% 的问题，这表明无需访问底层图像，基准中的相当一部分问题也能得到解答。这凸显了使用强大的大语言模型重新验证遥感任务的必要性，也说明需要构建真正要求进行视觉验证的基准数据集。

XLRS-Bench 上的结果。如表9所示，我们使用图像条件多模态模型和纯文本大语言模型基线（包括 GPT-4o [26], Qwen3-8B [73], 和 Llama3-8B [61]），在基于图像的视觉问答任务上评估 XLRS-Bench。在统一的评估协议下，纯文本大语言模型在推理任务上的表现出乎意料地具有竞争力：Qwen3-8B 的平均准确率达到**51。6**%，超过了图像条件 GPT-4o 基线（45。2%）。深入考察各子任务，Qwen3-8B 在72。0% 的 AD（异常检测）任务和 77。0% 的 ECR（存在性与计数推理）任务中取得佳绩；Llama3-8B 在 48。0% 的 RP （路径规划）任务中取得佳绩，尽管没有图像输入，其表现仍优于大多数视觉语言模型。这些结果表明，仅凭稳健的先验知识和语言线索，就能完成高级空间推理和计数任务，不过视觉信息对于其他任务仍至关重要。

表7。我们的基准中任务分类体系、标注来源和问答格式概览。L1/L2/L3分别表示粗粒度、中粒度和细粒度任务级别，属性。列出了每个L3任务的属性代码。我们还总结了每项任务的标注来源、问题生成器、问题类型、答案类型以及问题数量（#Q）。

L1任务	L2任务	L3任务	属性。	标注	问题生成器	问题类型	答案类型	#Q
感知	局部属性	颜色检测	COL	全人类	Qwen2.5-VL-7B	单张图像 单轮	多项选择（A/B/C/D） 开放式文本	443
		形状/边缘识别	SHP	全人类	Qwen2.5-VL-7B	单图单轮	多项选择（A/B/C/D） 开放式文本	243
		方向检测	DET	全人类	GPT-5 Thinking	单图单轮	多项选择（A/B/C/D） 开放式文本	228
	整体属性	物体分类	OC	全人类	Qwen2.5-VL-7B	单图单轮	多项选择（A/B/C/D） 开放式文本	436
		物体空间关系	REL	半自动化	Qwen2.5-VL-7B	单图单轮	多项选择（A/B/C/D） 开放式文本	444
	视觉定位	目标定位	OGD	全人类	GPT-5 Thinking	单图单轮	多项选择（A/B/C/D） 开放式文本	430
		区域定位	RG	全人类	GPT-5 Thinking	单图单轮	多项选择（A/B/C/D） 开放式文本	219
	计数	物体计数	OCN	半自动化	GPT-5 Thinking	单图单轮	多项选择（A/B/C/D） 开放式文本	412
		区域计数	RCN	全人类	Qwen2.5-VL-7B	单图单轮	多项选择（A/B/C/D） 开放式文本	245
推理	逻辑基础推理	多区域联合对比	MRJC MRJCS	全人类	GPT-5 Thinking	多图单轮 单图多框	选择题（A/B/C） 开放式文本	140
		目标状态判断	OSJ MOSJ	半自动化	GPT-5 Thinking	单图单轮 单图多轮	多项选择（A/B/C/D） 开放式文本	409
	知识支撑推理	异常检测	AD MAD	半自动化	GPT-5 Thinking	单图单轮 单图多轮	多项选择（A/B/C/D） 开放式文本	246
		未来预测	FP2I MTFP	半自动化	GPT-5 Thinking	单图单轮 单图多轮 多图单轮	多项选择（A/B/C/D） 多项选择（A/B） 开放式文本	429
图像描述	整体描述 区域描述		OCAP RCAP	半自动化 半自动化	GPT-5 Thinking GPT-5 Thinking	单图单轮 单图单轮	开放式文本 开放式文本	3913

表 8. MME-RealWorld 遥感子集上的准确率。所有模型均在仅使用文本输入（无图像）的情况下进行遥感任务评测。

方法	输入类型	感知（%）	#正确数 / #总数
Llama3-8B [45]	无图像	31.22	1167 / 3738
GPT-5-All [47]	无图像	22.45	839 / 3738
Gemini2.0-Flash [12]	无图像	16.19	605 / 3738
Qwen3-8B [73]	无图像	15.09	564 / 3738
GPT-4o [26]	无图像	2.03	76 / 3738

LRS-VQA 上的结果。如表10所示，我们在 LRS-VQA 上评估了多个大语言模型 [43]。相比之下，纯文本模型的表现明显较差：GPT-4（纯文本）的准确率为 19.71%，Qwen（纯文本）为 16.62%，而 GPT-4o-mini（纯文本）仅为 8.40%。Llama 3 8B（纯文本）几乎完全无法完成任务，在所有类别中的准确率都接近于零。这些结果表明，单靠语言先验无法解决 LRS-VQA，必须借助遥感图像中的视觉信息。

对输入分辨率的鲁棒性。 表 11报告了 XLRSBench 在 2K、4K 和 8K 分辨率下的结果。我们在相同设置下评估了 Qwen2.5-VL-7B、InternVL 3.5 和 Phi-3.5-Vision。对于这三个模型，准确率曲线几乎保持平坦：将分辨率从 8K 调整为 4K 或 2K，性能变化最多约为 2%。各模型的平均得分也呈现相同趋势（2K 下为 38.33%，8K 下为 39.75%）。这些结果表明，XLRSBench 的性能主要由与分辨率无关的语义线索决定，而不是由原始像素密度决定。

10. 案例研究与提示模板

各类任务的案例研究。 我们设计了一系列任务专用提示词，用于构建高质量视觉问答数据，系统覆盖遥感图像中的细粒度感知与推理。如图 1所示，我们从基础感知能力入手，包括在明确指定的参考区域内进行颜色检测、形状/边缘识别、方向检测和物体分类。图 2随后聚焦于以关系和定位为核心的感知任务，例如目标空间关系、目标依据、区域定位以及场景级物体计数。在此基础上，图 3 聚焦于区域感知，包括区域计数，以及在多图像和单图多边界框设置下的多区域联合比较，此时模型必须比较多个地块。图 4 介绍了物体状态判断提示词（单图、单轮和多轮）以及单图异常检测。图 5 进一步将异常推理扩展到多轮设置，并引入双图像未来预测。最后，图 6展示了单图多轮未来预测，要求模型首先定位关键工业综合体，然后根据周围的土地利用模式推理其潜在扩张情况。

面向细粒度感知与推理的提示设计。 为了细致分析模型行为，我们设计了特定任务提示模板，涵盖基准中的所有感知与推理类型。对于基础

表9。XLRS-Bench 在推理和感知维度上的结果。平均值表示该维度所有子任务的平均准确率。短横线（–）表示原始论文未报告相关结果。星号（*）表示在未输入图像的情况下进行评估的模型。

方法	推理						感知						
	AD	ECR	RP	RCCD	CCR	Avg.	OC	RC	RLUC	OCC	OCL	OSR	Avg.
遥感多模态大语言模型													
GeoChat [27]	33.0	43.0	10.0	-	21.0	26.8	16.7	29.0	23.0	21.1	16.8	24.2	21.8
GeoLLaVA-8K [62]	67.0	72.0	67.0	28.3	21.0	51.1	16.7	29.0	66.0	37.4	28.5	35.4	35.5
EarthDial [55]	62.0	71.0	43.0	48.3	50.0	54.9	18.3	42.0	36.0	31.3	31.0	24.8	30.6
VHM [49]	42.0	53.0	46.0	28.3	21.0	38.1	16.7	30.0	26.0	21.4	16.8	25.6	22.8
闭源多模态大语言模型													
GPT-4o [26]	73.0	73.0	35.0	20.0	25.0	45.2	25.0	32.0	66.0	9.5	11.3	24.6	28.1
GPT-4o-mini [26]	71.0	71.0	29.0	6.7	30.0	41.5	23.3	25.0	59.5	40.9	31.0	23.6	33.9
开源多模态大语言模型													
InternVL3-8B [69]	77.0	82.0	36.0	21.7	50.0	53.3	40.0	39.0	71.5	44.5	30.8	25.2	41.8
Qwen2.5-VL-7B [5]	68.0	72.0	27.0	38.3	45.0	50.1	33.3	40.0	77.0	40.6	40.5	36.2	44.6
InternVL3-78B [69]	76.0	81.0	40.0	45.0	42.0	56.8	23.3	49.0	74.0	42.5	37.4	30.0	42.7
大语言模型（纯文本）													
Llama3-8B [61]	58.0	66.0	48.0	28.3	20.0	44.1	16.8	30.0	23.0	16.7	21.4	25.8	22.3
Qwen3-8B [73]	72.0	77.0	43.0	35.0	31.0	51.6	30.5	28.0	40.5	30.0	29.0	25.4	30.6
GPT-4o [26]	74.0	75.0	55.0	35.0	24.0	52.6	21.9	28.0	30.0	13.3	32.1	30.4	26.0
GPT-4o-mini [26]	73.0	79.0	41.0	30.0	29.0	50.4	30.5	32.0	27.5	16.7	36.1	26.0	28.1
Claude3.7-Sonet [2]	62.0	74.0	45.0	18.3	22.0	44.3	18.9	29.0	28.0	16.7	30.4	25.4	24.7
视觉语言模型*（仅文本）													
GeoLLaVA-8K* [62]	63.0	71.0	63.0	25.0	26.0	49.6	23.5	24.0	48.0	16.7	25.1	35.6	28.8
EarthDial* [55]	58.0	74.0	33.0	55.0	53.0	54.6	31.1	37.0	27.0	23.3	35.6	24.4	29.7
VHM* [49]	35.0	51.0	46.0	28.3	21.0	36.3	16.8	29.0	25.0	16.7	21.2	25.2	22.3
GeoChat* [27]	44.0	52.0	46.0	28.3	23.0	38.7	16.8	29.0	23.0	16.7	21.1	29.0	22.6

表10. 通用视觉语言模型在 LRS-VQA 上的各类别准确率（%）。RU：农村/城市；ObjStatus：物体状态；ObjCat：物体类别；ObjBg：物体背景。

方法	RU	计数	物体状态	原因	物体类别	物体形状	物体颜色	物体背景	总体
GPT-4 [26]	32.11	0.00	13.70	12.60	12.77	44.63	30.98	8.98	19.71
Qwen3-8B [73]	37.94	0.17	5.10	9.30	10.23	37.40	19.61	6.53	16.62
GPT-4o-mini [26]	29.15	0.00	0.40	3.40	5.78	15.14	1.31	4.90	8.40

表11. XLSBench 在不同分辨率下的准确率。行表示模型，列表示输入分辨率。

模型	2K	4K	8K	Avg
QwenVL2.5-7B [5]	39.51	42.19	41.49	41.06
InternVL 3.5 [69]	38.41	38.60	39.84	38.95
Phi3.5-Vision [1]	37.08	37.11	37.92	37.37
Avg	38.33	39.30	39.75	39.13

感知，我们提供了颜色检测、形状/边缘识别、方向估计和目标分类的提示词（图 7–10），以及相对空间关系建模、目标定位、区域定位、目标级和区域级计数的提示词（图 11–15）。除单区域感知外，我们还在多图像和单图设置下引入了多区域联合对比提示词（图16 和 17）。针对更高层次的状态推理，我们构建了目标状态判断和异常检测提示模板，涵盖单轮和多轮对话形式（图19、 18、 20和21）。最后，我们为单图多轮对话和双图时间对比场景设计了未来预测提示词（图22 和 23）。这些统一的提示模板.....

直观展示模型如何处理多样化的遥感任务，并通过系统的定性分析补充定量评估。

视觉问答分数的提示模板。图 24展示了我们为不提供答案选项的开放式视觉问答设计的评分提示词。该提示词将 GPT-5-Thinking 设定为专家评审，要求其给出 1–100 分的评分；评分优先考虑答案相对于参考答案的正确性，同时也考虑实用性、完整性和幻觉惩罚。提示词列出了宽松但明确的匹配准则，涵盖语义等价、数值和范围容差、单位处理、方向性术语，以及对“是/否”和无法作答情形的特殊处理，从而避免对无害的改述或细微偏差过度扣分。此外，我们还提供了建议的分数区间（例如，语义等价且准确的答案为 80–100 分），以校准评审并促使其一致地使用完整评分范围。最后，模板将输出限定为恰好两行，且仅包含数字分数，从而便于可靠地自动解析和汇总视觉问答分数。

图像描述提示模板。图25和26展示了我们为 GPT-5-Thinking 设计的提示词。第一个提示词引导模型为高分辨率航空或卫星图像生成全局场景描述，包括整体摘要和象限级子区域分析。第二个提示词将模型转变为按象限组织描述的高分辨率遥感图像描述器，要求提供全局摘要，并明确列出自然要素和人工要素的使用情况，同时施加精确的空间与功能约束。

确列出自然要素和人工要素的使用情况，并施加精确的空间与功能约束。

参考文献

[1] Marah Abdin, Jyoti Aneja, Harkirat Behl, S´ebastien Bubeck, Ronen Eldan, Suriya Gunasekar, Michael Harrison, Russell J Hewett, Mojan Javaheripi, Piero Kauffmann 等。Phi-4 技术报告。arXiv 预印本 arXiv:2412.08905，2024。1, 7, 8, 10, 12[2] AI Anthropic。Claude 3.5 sonnet 模型卡增补。Claude-3.5 模型卡，3(6)，2024。1, 10, 12[3] 多位作者。MME：多模态大语言模型综合评估基准。arXiv 预印本 arXiv:2406.00381，2024。2[4] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou 和 Jingren Zhou。Qwen-vl：一种用于理解、定位、文本阅读等任务的通用型视觉语言模型。arXiv 预印本 arXiv:2308.12966，2023。3[5] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu 和 Junyang Lin。Qwen2.5-vl 技术报告。arXiv 预印本 arXiv:2502.13923，2025。1, 2, 4, 7, 8, 10, 12[6] John E Ball, Derek T Anderson 和 Chee Seng Chan。深度学习在遥感领域的综合综述：理论、工具及面向社区的挑战。《应用遥感期刊》，11(4):042609–042609，2017。1[7] Yakoub Bazi, Laila Bashmal, Mohamad Mahmoud Al Rahhal, Riccardo Ricci 和 Farid Melgani。RS-LLaVA：一种用于遥感图像联合图像描述和问题回答的大型视觉语言模型。《遥感》，16(9)，2024。3[8] Wouter Bulten, Kimmo Kartasalo, Po-Hsuan Cameron Chen 等。用于前列腺癌诊断和格里森分级的人工智能：PANDA 挑战赛。《自然医学》，28(1):154–163，2022。2, 3[9]Javiera Castillo-Navarro, Bertrand Le Saux, Alexandre Boulch, Nicolas Audebert 和 S´ebastien Lef`evre。地球观测中的半监督语义分割：MiniFrance 套件、数据集分析与多任务网络研究。《机器学习》，111(9):3125–3160，2022。2, 3[10] Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollar 和 C. Lawrence Zitnick。Microsoft COCO Captions：数据收集与评估服务器。arXiv 预印本 arXiv:1504.00325，2015。2[11] Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, Bin Li, Ping Luo, Tong Lu, Yu Qiao 和 Jifeng Dai。InternVL：扩展视觉基础模型并使其适配通用视觉语言任务。arXiv 预印本 arXiv:2312.14238，2023。1, 7, 8[12] Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, 等。Gemini 2.5：凭借先进推理、多模态、长上下文和新一代智能体能力，推动前沿发展。arXiv 预印本 arXiv:2507.06261，2025。7, 8，11[13]Muhammad Danish、Muhammad Akhtar Munir、Syed Roshaan Ali Shah、Kartik Kuckreja、Fahad Shahbaz Khan、Paolo Fraccaro、Alexandre Lacoste 和 Salman Khan。Geobench-vlm：面向地理空间任务的视觉语言模型基准测试。载于 IEEE/CVF计算机视觉国际会议论文集，第 7132–7142 页，2025。1 – 。[14] Rodrigo Caye Daudt、Bertrand Le Saux、Alexandre Boulch 和 Yann Gousseau。面向大规模语义变化检测的多任务学习。《计算机视觉与图像理解》，187:102783，2019。2，3[15] Jian Ding、Nan Xue、Gui-Song Xia、Xiang Bai、Wen Yang、Michael Yang、Serge Belongie、Jiebo Luo、Mihai Datcu、Marcello Pelizzola 和 Liang-Pei Zhang。DOTA v2.0：航空影像目标检测数据集。《IEEE地球科学与遥感汇刊》，60（15）：1–12，2022。1，2，3，5。[16] Team GLM 等。ChatGLM：从 GLM-130B 到 GLM-4 All Tools 的大型语言模型系列，2024。10[17] Zhang 等。HRSCD：高分辨率场景分类数据集。IEEE计算机视觉与模式识别会议（CVPR）论文集，2022。3[18] Yash Goyal、Tejas Khot、Douglas Summers-Stay、Dhruv Batra 和 Devi Parikh。让 VQA 中的 V 发挥作用：重新审视视觉问答。IEEE计算机视觉与模式识别会议（CVPR）论文集，2017。2[19] Zonghao Guo、Ruyi Xu、Yuan Yao、Junbo Cui、Zanlin Ni、Chunjiang Ge、Tat-Seng Chua、Zhiyuan Liu 和 Gao Huang。LLaVA-UHD：可感知任意宽高比和高分辨率图像的大型多模态模型。欧洲计算机视觉会议（ECCV）论文集，2024。1，3[20] Danna Gurari、Qing Li、Abigale J. Stangl、Anhong Guo、Chi Lin、Kristen Grauman、Jiebo Luo 和 Jeffrey P. Bigham。VizWiz大挑战：回答盲人的视觉问题。IEEE计算机视觉与模式识别会议（CVPR）论文集，2018。2[21] Qiu Han、Gongjie Zhang、Jiaxing Huang、Peng Gao、Zhang Wei 和 Shijian Lu。面向大规模视觉变换器的高效掩码自编码器。IEEE/CVF冬季应用计算机视觉会议（WACV）论文集，第 606–615 页，2024。3[22] Junyuan Hong、Jinhao Duan、Chenhui Zhang、Zhangheng Li、Chulun Xie、Kelsey Lieberman、James Diffenderfer、Brian Bartoldson、Ajay Jaiswal、Kaidi Xu、Bhavya Kailkhura、Dan Hendrycks、Dawn Song、Zhangyang Wang 和 Bo Li。解码压缩环境下的可信度：审视压缩下高效大语言模型的可信性。arXiv 预印本 arXiv:2403.15447，2024。3[23] Jingliang Hu、Rong Liu、Danfeng Hong、Andr´es Camero、Jing Yao、Mathias Schneider、Franz Kurz、Karl Segl 和 Xiao Xiang Zhu。MDAS：面向遥感的新型多模态基准数据集。《地球系统科学数据》，15（1）：113–131，2023。3

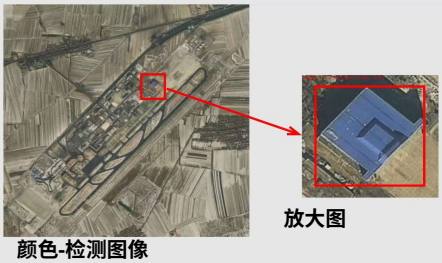
[24] Liang Huang、Fengxiang Wang、Yalun Zhang 和 Qingxia Xu。结合卷积神经网络与 Swin Transformer 的细粒度船舶分类。《遥感》，14(13):3087，2022。3[25] Drew A. Hudson 和 Christopher D. Manning。GQA：面向真实世界视觉推理和组合式问答的新数据集。IEEE计算机视觉与模式识别会议（CVPR）论文集，2019。2[26] Aaron Hurst、Adam Lerer、Adam P Goucher、Adam Perelman、Aditya Ramesh、Aidan Clark、AJ Ostrow、Akila Welihinda、Alan Hayes、Alec Radford 等。GPT-4o 系统卡。arXiv 预印本 arXiv:2410.21276，2024。1，3，6，7，8，10，11，12[27] Kartik Kuckreja、Muhammad Sohail Danish、Muzammal Naseer、Abhijit Das、Salman Khan 和 Fahad Shahbaz Khan。GeoChat：面向遥感的基于视觉定位的大型视觉语言模型。IEEE/CVF计算机视觉与模式识别会议论文集，第 27831–27840 页，2024。1，3，7，8，10，12[28] Kun Li、George Vosselman 和 Michael Ying Yang。HRVQA：面向高分辨率航空影像的视觉问答基准。《ISPRS摄影测量与遥感杂志》，214:65–81，2024。2，5[29] Wei Li、Peng Li、Xue Li 等。DIOR：光学遥感图像目标检测的大规模基准。《IEEE地球科学与遥感汇刊》，59（12）：1–15，2021。5[30] Wei Li、Yao Sun、Peng Li 等。MillionAID：面向场景分类的百万级航空影像数据集。《IEEE地球科学与遥感汇刊》，60(8):1–12，2022。5[31] Xiang Li、Jian Ding 和 Mohamed Elhoseiny。VRSBench：用于遥感图像理解的通用视觉语言基准数据集。arXiv 预印本 arXiv:2406 12384，2024。1，2，3，5，7。[32] Yan wei Li、Yuechen Zhang、Chengyao Wang、Zhisheng Zhong、Yixin Chen、Ruihang Chu、Shaoteng Liu 和 Jiaya Jia。Mini-Gemini：挖掘多模态视觉语言模型的潜力。arXiv 预印本 arXiv:2403 18814，2024。1，3。[33] Zhang Li、Biao Yang、Qiang Liu、Zhiyin Ma、Shuo Zhang、Jingxu Yang、Yabo Sun、Yuliang Liu 和 Xiang Bai。Monkey：图像分辨率和文本标签对大型多模态模型至关重要。IEEE/CVF计算机视觉与模式识别会议（CVPR）论文集，2024。3[34] Haotian Liu、Chunyuan Li、Yuheng Li 和 Yong Jae Lee。视觉指令微调。arXiv 预印本 arXiv:2304 08485，2023。3。[35] Haotian Liu、Chunyuan Li、Yuheng Li、Bo Li、Yuanhan Zhang、Sheng Shen 和 Yong Jae Lee。LLaVA-NeXT：改进推理、光学字符识别和世界知识。arXiv 预印本 arXiv:2310.03744，2023。1，3[36] Mingyu Liu、Ekim Yurtsever、Jonathan Fossaert、Xingcheng Zhou、Walter Zimmer、Yuning Cui、Bare Luka Zagar 和 Alois C Knoll。自动驾驶数据集综述：统计数据、标注质量与未来展望。《IEEE智能车辆汇刊》，2024。1

[37] Yuan Liu、Haodong Duan、Yuanhan Zhang、Bo Li、Songyang Zhang、Wangbo Zhao、Yike Yuan、Jiaqi Wang、Conghui He、Ziwei Liu、Kai Chen 和 Dahua Lin。MMBench：你的多模态模型是全能选手吗? arXiv 预印本 arXiv:2307.06281，2023。2[38] Sylvain Lobry、Diego Marcos、Jesse Murray 和 Devis Tuia。RSVQA：遥感数据的视觉问答。IEEE地球科学与遥感汇刊，58（12）：8555–8566，2020。2，5，6[39]Haoyu Lu、Wen Liu、Bo Zhang、Bingxuan Wang、Kai Dong、Bo Liu、Jingxiang Sun、Tongzheng Ren、Zhuoshu Li、Hao Yang 等。DeepSeek-VL：迈向真实场景中的视觉语言理解。arXiv 预印本 arXiv:2403 05525，2024 7，8，10 . . [40] Xiaoqiang Lu、Binqiang Wang、Xiangtao Zheng 和 Xue-long Li。探索遥感图像描述生成的模型与数据。IEEE地球科学与遥感汇刊，56(11)：2183–2195，2017。5[41] Xiaoqiang Lu、Binqiang Wang、Xiangtao Zheng 和 Xue-long Li。RSICap：高质量遥感图像描述数据集。IEEE地球科学与遥感汇刊，61(5)：1–14，2023。5[42] Junwei Luo、Zhen Pang、Yongjun Zhang、Tingzhu Wang、Linlin Wang、Bo Dang、Jiangwei Lao、Jian Wang、Jingdong Chen、Yihua Tan 和 Yansheng Li。SkySenseGPT：用于遥感视觉语言理解的细粒度指令微调数据集与模型。arXiv 预印本 arXiv:2406.10100，2024。3 [43] Junwei Luo、Yingying Zhang、Xue Yang、Kang Wu、Qi Zhu、Lei Liang、Jingdong Chen 和 Yansheng Li。当大型视觉语言模型遇上大型遥感图像：由粗到细的文本引导令牌剪枝。arXiv 预印本 arXiv:2503.07588，2025。2，5，6，11[44] Kenneth Marino、Mohammad Rastegar、Ali Farhadi 和 Roozbeh Mottaghi。OK-VQA：需要外部知识的视觉问答基准。IEEE计算机视觉与模式识别会议（CVPR）论文集，2019。2[45] Meta AI。Llama 3——模型卡与提示词格式，2024。Llama 3 系列的模型卡。4，7，8，11[46] Dilxat Muhtar、Zhenshi Li、Feng Gu、Xueliang Zhang 和 Pengfeng Xiao。LHRS-Bot：借助志愿地理信息增强的大型多模态语言模型赋能遥感。载于欧洲计算机视觉会议，第440–457页。Springer，2024。1，3[47] OpenAI。GPT-5系统卡。技术报告，OpenAI，2025。介绍 gpt-5-main 和 gpt-5-thinking 的系统卡。1，2，4，6，7，8，11[48] Chao Pang、Xingxing Weng、Jiang Wu、Jiayu Li、Yi Liu、Jiaxing Sun、Weijia Li、Shuai Wang、Litong Feng、Gui-Song Xia 和 Conghui He。H2RSVLM：迈向有帮助且诚实的遥感大型视觉语言模型。arXiv 预印本，2024。3[49] Chao Pang、Xingxing Weng、Jiang Wu、Jiayu Li、Yi Liu、Jiaxing Sun、Weijia Li、Shuai Wang、Litong Feng、Gui-Song Xia 等。VHM：用于遥感图像分析的通用且诚实的视觉语言模型。论文发表于

AAAI人工智能会议，39(6)：6381–6388，2025。1，3，7，8，10，12[50] Bryan A. Plummer、Liwei Wang、Chris M. Cervantes、Juan C. Caicedo、Julia Hockenmaier 和 Svetlana Lazebnik。Flickr30k实体：收集区域与短语的对应关系，以构建更丰富的图像到句子模型。IEEE国际计算机视觉会议（ICCV）论文集，2015。2[51] Yangjun Ruan、Yuanhao Xiong、Sashank Reddi、Sanjiv Kumar 和 Cho-Jui Hsieh。通过零阶预言机学习如何学习。国际机器学习会议（ICML）论文集，2020。3[52]Akashah Shabbir、Mohammed Zumri、Mohammed Bennamoun、Fahad S Khan 和 Salman Khan。GeoPixel：在遥感中进行像素定位的大型多模态模型。arXiv 预印本 arXiv:2501.13925，2025。5[53] Baifeng Shi、Boyi Li、Han Cai、Yao Lu、Sifei Liu、Marco Pavone、Jan Kautz、Song Han、Trevor Darrell、Pavlo Molchanov 等。将视觉预训练扩展至 4K 分辨率。arXiv 预印本 arXiv:2503.19903，2025。1，7，8，10[54] Amanpreet Singh、Vivek Natarjan、Meet Shah、Yu Jiang、Xinlei Chen、Dhruv Batra、Devi Parikh 和 Marcus Rohrbach。TextVQA：文本视觉问答数据集。IEEE计算机视觉与模式识别会议（CVPR）论文集，2019。2[55] Sagar Soni、Akshay Dudhane、Hiyam Debary、Mustansar Fiaz、Muhammad Akhtar Munir、Muhammad Sohail Danish、Paolo Fraccaro、Campbell D Watson、Levente J Klein、Fahad Shahbaz Khan 等。EarthDial：将多传感器地球观测转化为交互式对话。计算机视觉与模式识别会议论文集，第14303–14313页，2025。1，7，8，10，12[56] Xian Sun、Peijin Wang、Zhiyuan Yan、Feng Xu、Ruiping Wang、Wenhui Diao、Jin Chen、Jihao Li、Yingchao Feng、Tao Xu、Martin Weinmann、Stefan Hinz、Cheng Wang 和 Kun Fu。FAIR1M：用于高分辨率遥感图像细粒度目标识别的基准数据集。《摄影测量与遥感杂志》，184：116–130，2022。2，3，5[57] Gemini团队和 Google。Gemini：一系列能力强大的多模态模型。arXiv 预印本 arXiv:2312 11805，2023 3 . . [58] T.K.K.Borgeaud、A.Liu、S.M.Almeida 等。Llama：开放且高效的基础语言模型。arXiv 预印本 arXiv:2302.13971，2023。3[59]Shengao Tong、Ellis Brown、Penghao Wu、Sanghyun Woo、Manoj Middepogu、Sai Charitha Akula、Jihan Yang、Shusheng Yang、Adithya Iyer、Xichen Pan 等。Cambrian-1：以视觉为中心的多模态大语言模型开放式探索。arXiv 预印本 arXiv:2406.16860，2024。3[60]Lucrezia Tosato、Hichem Boussaud、Flora Weissgerber、Camille Kurtz、Laurent Wendling 和 Sylvain Lobry。用于遥感图像视觉问答的分割引导注意力。IEEE国际地球科学与遥感研讨会（IGARSS）论文集，第2750–2754页，2024。5[61]Hugo Touvron、Thibaut Lavril、Gautier Izacard、Xavier Martinet、Marie-Anne Lachaux、Timoth´ee Lacroix、Baptiste Rozi`ere, Naman Goyal, Eric Hambro, Faisal Azhar, 等。Llama：开放且高效的基础语言模型。arXiv 预印本 arXiv:2302.13971，2023。10, 12[62] Fengxiang Wang, Mingshuo Chen, Yueying Li, Di Wang, Haotian Wang, Zonghao Guo, Zefan Wang, Boqi Shan, Long Lan, Yulin Wang, Hongzhen Wang, Wenjing Yang, Boqi Du, 和 Jing Zhang。GeoLLaVA-8K：将遥感多模态大语言模型扩展至 8k 分辨率。arXiv 预印本 arXiv:2505.21375，2025。1, 3, 7, 10, 12[63] Fengxiang Wang, Hongzhen Wang, Mingshuo Chen, Di Wang, Yulin Wang, Zonghao Guo, Qiang Ma, Long Lan, Wenjing Yang, Jing Zhang, Zhiyuan Liu, 和 Maosong Sun。XLRS-Bench：你的多模态大语言模型能否理解超大幅超高分辨率遥感图像？载于 IEEE/CVF计算机视觉与模式识别会议（CVPR）论文集，第 14325–14336 页，2025。1, 2, 3, 5, 6, 7[64] Fengxiang Wang, Hongzhen Wang, Zonghao Guo, Di Wang, Yulin Wang, Mingshuo Chen, Qiang Ma, Long Lan, Wenjing Yang, 和 Jing Zhang。RSIEval：用于评估遥感图像描述的基准。IEEE/CVF计算机视觉与模式识别会议（CVPR）论文集，第 14325–14336 页，2025。2, 3, 5[65] Fengxiang Wang, Hongzhen Wang, Yulin Wang, Di Wang, Mingshuo Chen, Haiyan Zhao, Yangang Sun, Shuo Wang, Long Lan, Wenjing Yang, 和 Jing Zhang。Roma：扩展用于遥感的基于 Mamba 的基础模型。arXiv 预印本 arXiv:2503.10392，2025。3[66] Haochen Wang, Xiangtai Li, Zilong Huang, Anran Wang, Jiacong Wang, Tao Zhang, Jiani Zheng, Sule Bai, Zijian Kang, Jiashi Feng, 等。增强视觉基础推理中的可追溯证据：评估与方法。arXiv 预印本 arXiv:2507.07999，2025。1, 2[67] Junjue Wang, Zhuo Zheng, Zihang Chen, Ailong Ma, 和 Yanfei Zhong。EarthVQA：基于关系推理的遥感视觉问答，迈向可查询地球。AAAI人工智能会议论文集，38(6):5481–5489，2024。5[68] Wenbin Wang, Liang Ding, Minyan Zeng, Xiabin Zhou, Li Shen, Yong Luo, Wei Yu, 和 Dacheng Tao。分而治之，再合并：一种用于多模态大语言模型高分辨率图像感知的免训练框架。载于 AAAI人工智能会议论文集，第 7907–7915 页，2025。2[69] Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, 等。InternVL3.5：提升开源多模态模型的通用性、推理能力与效率。arXiv 预印本 arXiv:2508.18265，2025。1, 7, 8, 10, 12[70] Penghao Wu 和 Saining Xie。V?：将引导式视觉搜索作为多模态大语言模型的核心机制。载于 IEEE/CVF计算机视觉与模式识别会议论文集，第 13084–13094 页，2024。1[71] Gui-Song Xia, Xiang Bai, Jian Ding, Zhen Zhu, Serge Belongie, Jiebo Luo, Mihai Datcu, Marcello Pelillo, 和 Liangpei Zhang。DOTA：用于航空影像目标检测的大规模数据集。IEEE计算机视觉与模式识别会议论文集，第

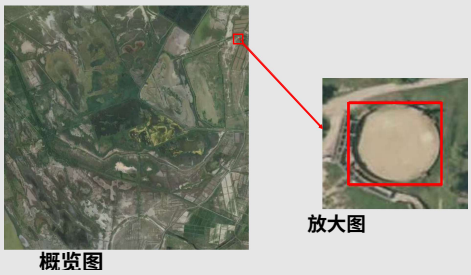
3974 3983 页，2018 2, 3, 5 – 。[72] Jianwen Xie, Yichao Lu, Yuxin Peng, 等。Seedbench：用于评估多模态大语言模型在真实场景中表现的基准。IEEE国际计算机视觉会议（ICCV）论文集，2023。2[73]An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, 等。Qwen3 技术报告。arXiv 预印本 arXiv:2505.09388，2025。4, 7, 8, 10, 11, 12[74] Kelu Yao, Nuo Xu, Rong Yang, Yingying Xu, Zhuoyan Gao, Titinunt Kitrungrotsakul, Yi Ren, Pu Zhang, Jin Wang, Ning Wei, 等。Falcon：遥感视觉语言基础模型。arXiv 预印本 arXiv:2503.11070，2025。3[75] Yuan Yao, Tianyu Yu, Ao Zhang, Chongyi Wang, Junbo Cui, Hongji Zhu, Tianchi Cai, Haoyu Li, Weilin Zhao, Zhihui He, 等。MiniCPM-V：媲美 GPT-4V 水平、可在手机上运行的多模态大语言模型。arXiv 预印本 arXiv:2408.01800，2024。1, 3, 7, 8, 10[76] Kejia Yin, Varshanth Rao, Ruowei Jiang, Xudong Liu, Parham Aarabi, 和 David B. Lindell。SCE-MAE：利用掩码自编码器选择性增强对应关系，以进行自监督地标估计。IEEE/CVF计算机视觉与模式识别会议（CVPR）论文集，第1313–1322 页，2024。3[77] Kaining Ying, Fanqing Meng, Jin Wang, Zhiqian Li, Han Lin, Yue Yang, Hao Zhang, Wenbo Zhang, Yuqi Lin, Shuo Liu, 等。MMT-Bench：面向多任务通用人工智能的大型视觉语言模型评估综合多模态基准。arXiv 预印本 arXiv:2404.16006，2024。2[78] Zhenghang Yuan, Lichao Mou, Yuansheng Hua, 和 Xiao Xiang Zhu。RRSIS：指代遥感图像分割。IEEE/CVF计算机视觉与模式识别会议（CVPR）论文集，第26658–26668 页，2024。5[79] Yang Zhan, Zhitong Xiong, 和 Yuan Yuan。RSVG：探索遥感数据上的视觉定位数据与模型。IEEE地球科学与遥感汇刊，61: 1–13，2023。5[80] Yang Zhan, Zhitong Xiong, 和 Yuan Yuan。SkyEyeGPT：通过基于大语言模型的指令微调，统一遥感视觉语言任务。ISPRS摄影测量与遥感杂志，221:64–77，2025。早期版本：arXiv:2401.09712。3[81] Cheng-Fan Zhang, Haochen Tian, Chaoyou Fu, Shuangqing Zhang, Junfei Wu, Feng Li, Kun Wang, Qingsong Wen, Zhang Zhang, Liang Wang, Rong Jin, 和 Tieniu Tan。擅长图像描述，却不擅长计数：基准测试 GPT-4V 在地球观测数据上的表现。arXiv 预印本 arXiv:2401 17600，2024 3 . .[82] Pan Zhang, Xiaoyi Dong, Bin Wang, Yuhang Cao, Chao Xu, Linke Ouyang, Zhiyuan Zhao, Shuangrui Ding, Songyang Zhang, Haodong Duan, Hang Yan, Xinyue Zhang, Wei Li, Jingwen Li, Kai Chen, Conghui He, Xingcheng Zhang, Yu Qiao, Dahua Lin, 和 Jiaqi Wang。InternLM-XComposer：用于高级图文理解与创作的视觉语言大模型。arXiv 预印本 arXiv:2309.15112，2023。3[83] Wei Zhang、Cheng Liu、Yao Sun、Peng Li 和 Xin Zhao。re-fgeo：一个大规模指代表达遥感图像数据集。IEEE/CVF计算机视觉与模式识别会议论文集（CVPR），第12345–12355页，2023年。5 – . [84] Wei Zhang、Miaoxin Cai、Tong Zhang、Yin Zhuang 和 Xuerui Mao。EarthGPT：面向遥感领域多传感器图像理解的通用多模态大语言模型。IEEE地球科学与遥感汇刊，62，2024年。3[85] X. Zhang、H. Boussaid、L. Tosato、F. Weissgerber、C. Kurtz、L. Wendling 和 S. Lobry。多种遥感图像模态上的视觉问答。arXiv预印本 arXiv:2505.15401，2025年。5[86] Yi-Fan Zhang、Huanyu Zhang、Haochen Tian、Chaoyou Fu、Shuangqing Zhang、Junfei Wu、Feng Li、Kun Wang、Qingsong Wen、Zhang Zhang、Liang Wang、Rong Jin 和 Tieniu Tan。MME-RealWorld：你们的多模态大语言模型能否挑战对人类而言困难的高分辨率真实场景? OpenReview预印本（arXiv/OpenReview），2025年。1, 2[87] Yi-Fan Zhang、Qingsong Wen、Chaoyou Fu、Xue Wang、Zhang Zhang、Liang Wang 和 Rong Jin。超越Llava-HD：深入探索高分辨率大型多模态模型。arXiv预印本 arXiv:2406.08487，2024年。1[88] Zilun Zhang、Haozhan Shen、Tiancheng Zhao、Bin Chen、Zian Guan、Yuhao Wang、Xu Jia、Yuxiang Cai、Yongheng Shang 和 Jianwei Yin。GeoRSMLLM：用于地球科学与遥感视觉语言任务的多模态大语言模型。arXiv预印本 arXiv:2503.12490，2025年。3[89] Yuanxin Zhao、Mi Zhang、Bingnan Yang、Zhan Zhang、Jujia Kang 和 Jianya Gong。LuoJiaHOG：面向遥感图像文本检索的层次化地理感知图像描述数据集。ISPRS摄影测量与遥感学报，222：130–151，2025年。5[90] Xiangtao Zheng、Binqiang Wang、Xingqian Du 和 Xiaoqiang Lu。用于遥感视觉问答的互注意力Inception网络。IEEE地球科学与遥感汇刊，60：1–14，2021年。5[91] Luowei Zhou、Chenliang Xu 和 Jason J. Corso。迈向从网络教学视频中自动学习流程。AAAI人工智能会议论文集，第7590–7598页，2018年。2[92] Deyao Zhu、Jun Chen、Xiaoqian Shen、Xiang Li 和 Mohamed Elhoseiny。MiniGPT-4：利用先进的大语言模型增强视觉语言理解。第十二届国际学习表征会议，2024年。3[93] Xing Zi、Jinghao Xiao、Yunxiao Shi、Xian Tao、Jun Li、Ali Braytee 和 Mukesh Prasad。RSVLM-QA：一个基于遥感视觉语言模型的问答基准数据集。arXiv预印本 arXiv:2508 07918，2025年。5 . .

类别：颜色检测
图像分辨率： 4701 x4946
问题： 确定图像中给定参考边界框内机场的颜色。边界框：[2814, 1628, 2975, 1790]。
选项：
A：棕色
B：蓝色
C：白色
D：灰色
正确答案： B

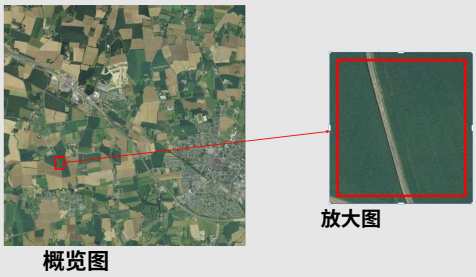


类别：形状/边缘识别 图像分辨率： 10000x10000**问题：** 图像中给定参考边界框内，主体对象边界的最精确特征/形状描述是以下哪一项？边界框：[9541, 1273, 9610, 1344]。

选项：
A：不规则
B：圆形C：椭圆形 D：正方形
正确答案： B



类别：方向检测
图像分辨率： 10000x10000
问题： 确定方向图像中给定参考边界框内道路网络的方向。边界框：[2096, 6332, 2454, 6732]。
选项：
A：右 B：上
C：右上方
D：右下
正确答案： D



类别：物体分类
图像分辨率： 10000x10000
问题： 确定类别，位于图像中给定参考边界框内的主体对象。边界框：[2382, 5403, 4489, 6853]。
选项：
A：沿海湿地 B：稀疏农田区域C：城市住宅区 D：茂密森林
正确答案： B



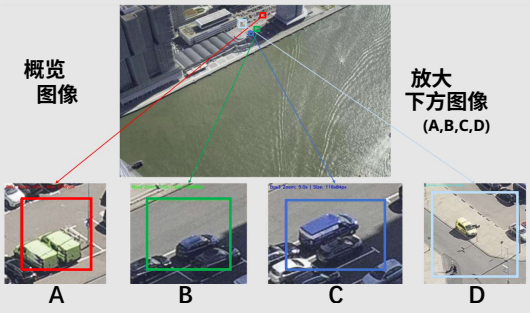
图 1. 案　例 f 来自不同任务：颜色检测、形状/边缘识别、方向检测和分类。

类别：物体空间关系
图像分辨率： 4183 x8115
问题： 在图像中，椭圆形体育场（带有跑道）位于何处相对于带有白色屋顶的矩形灰色建筑?
选项：
A：向左B：向右 C：下方 D：上方
正确答案： A



类别：目标定位
图像分辨率： 7360 x4912

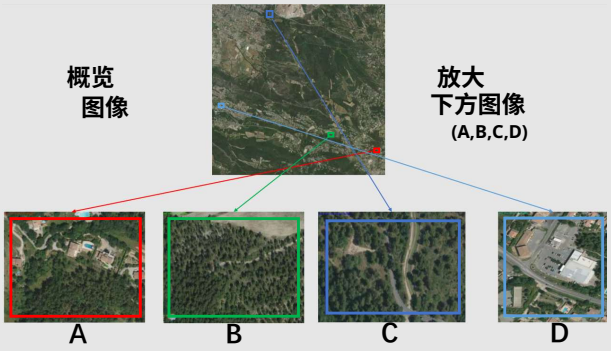
问题： 哪个边界框 最能定位 一辆绿色面包车?
选项：
A： [3806, 270, 3893, 362]
B： [3662, 656, 3761, 742]
C： [3500, 758, 3616, 842]
D： [3237, 418, 3430, 619]
正确答案： A



类别：区域定位
图像分辨率： 10000x10000

问题： 哪个边界框 最准确地定位由道路和停车场为主、周围环绕着植被和住宅区的商业综合体区域?

选项：
A: [9072, 8359, 9467, 8637]
B: [6514, 7541, 7036, 7805]
C: [3102, 436, 3480, 640]
D： [336, 5876, 636, 6271]
正确答案： D



类别：物体计数 图像分辨率： 7360x4912**问题：** 整个图像中有多少个 篮球场？ **选项：**

A: 4 B: 3 C: 5 D: 1

正确答案： D

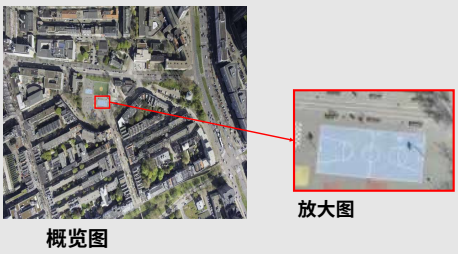
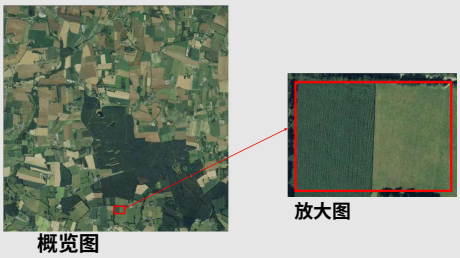


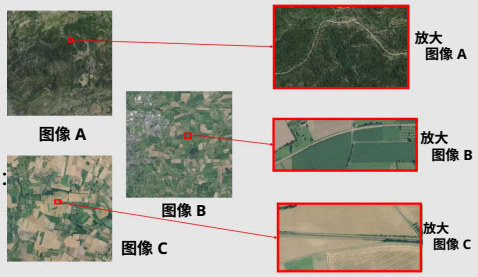
图2. 案例 fr　　来自不同任务：目标空间关系、目标定位、区域定位和物体计数　　.

类别：区域计数
图像分辨率： 10000x10000 **问题：**
有多少个陆地区块位于图像中给定的参考边界框内？边界框：
[4918, 8846, 5361, 9146]。
选项：
A:
1 B:
6 C:
2D:
4
正确答案： C



类别：多区域联合对比（多图像）
图像分辨率： 10000x10000

问题：
这三张图片中给出的三个边界框里，哪一个显示的是最蜿蜒的道路？图像 A 的边界框：[4446, 2282, 6175, 3339],图像 B 的边界框：[4404, 3882, 6046, 4453],图像 C 的边界框：[3825, 4061, 4911, 4546] （图像 A/B/C 根据其传输顺序确定。）分辨率：10000 x 10000，图像 A 的边界框：[4446, 2282, 6175, 3339],图像 B 的边界框：[4404,3882, 6046, 4453],图像 C 的边界框：[3825,4061, 4911, 4546]（图像 A/B/C 根据其传输顺序确定。）
选项：
A：图像 B B：无
法比较 C：图像
AD：图像 C
正确答案： C



类别：多区域联合对比（单图多框）
图像分辨率： 10000x10000

问题：
哪个边界框最能代表绿度最高的农田地块？
最高绿度？
选项：
A： [5054, 1461, 6196, 2453]
B： [7625, 5725, 8968, 6982]
C： [4175, 7411, 5654, 8632]
正确答案： A

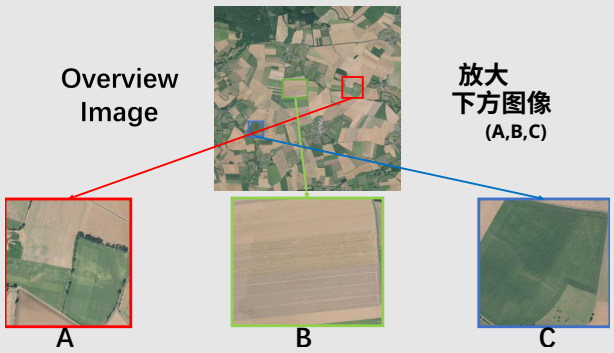


图3. 不同任务的案例：区域计数、多区域联合对比（多图像）、多区域联合对比（单图多框）。

类别：目标状态判断（单图多轮）
图像分辨率： 10000x10000

问题1:
图像左下方的建筑物有何特征？它被道路和田地组成的网格环绕。

选项:
A：屋顶覆盖着一层雪，表明处于冬季状况。B：屋顶状况良好，没有明显的损坏迹象。C：墙壁上有大面积裂缝，表明存在结构损坏。D：建筑物周围区域被淹，可以看到积水。

正确答案： B

问题2:
图像左下方建筑物的状况如何影响其运行状态?

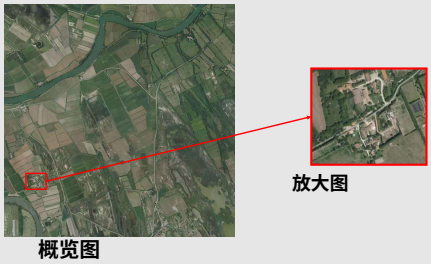
选项:
A：建筑物可能正在建设中，因为可以看到脚手架。B：建筑物似乎完全正常运行，没有任何明显的损坏迹象。C：由于结构中可见裂缝，该建筑物不适宜居住。D：建筑物被淹，无法正常使用。

正确答案： B

问题3:
图像左下角部分中建筑物的当前状态是什么?

选项:
A：完好B：受损 C：建设中 D：被淹

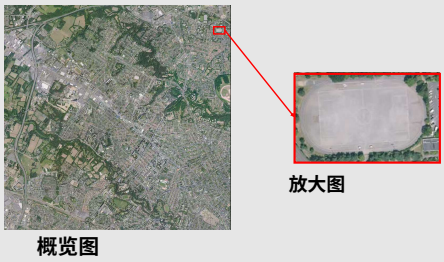
正确答案： A



类别：目标状态判断（单图单轮）
图像分辨率： 10000x10000 **问题:**
位于图像右上角、具有土跑道和绿色内场的大型矩形运动场，其当前状态是什么?

选项:
A：建设中 B：受损
C：未使用／非活跃
D：使用中／活跃

正确答案： D



类别：异常检测（单图单轮）
图像分辨率： 10000x10000

问题:
图像中间偏右区域有一块颜色较深的绿色大矩形田地。这块田地颜色较周围田地更深的最可能原因是什么?

选项:
A：这块田地最近施过肥，因此颜色较深。B：这块田地出现了渍水，导致颜色较深。C：这块田地正在进行轮作，因此颜色较深。D：这块田地用于放牧动物，因此颜色较深。

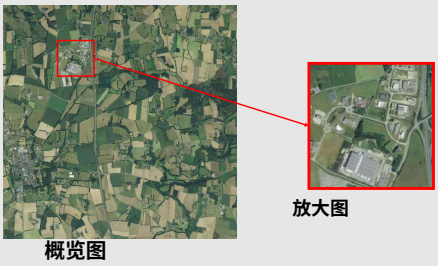
正确答案： B



图4. 不同任务的案例：目标状态判断（单图多轮）、目标状态判断（单图单轮）、异常检测（单图单轮）。

类别：异常检测（单图多轮）
图像分辨率： 10000x10000

问题1：
在哪里有一处矩形浅灰色工业园区，其中聚集着几座
大型建筑物和停车场，周围环绕着开阔的绿色和黄色田地
，并且毗邻一条南北向笔直的高速公路？它位于图像中的什么位置？**选项：**
A：图像的右上角。B：图像的中间。
C： 图像的左上角。D：图像的左下角。
正确答案： C



问题2：
根据图像的视觉线索，什么关系或变化可能影响所选对象/区域在不久的将来？
选项：
A：由于空间充足，工业综合体向附近农田扩张。B：洪涝风险导致工业活动减少。C：该区域转变为住宅开发项目。D：综合体被废弃，导致植被杂草丛生。

正确答案： A

问题3：
如果由于空间充足，工业综合体向附近农田扩张发生，那么所选物体/区域最可能的未来状态是什么？

选项：
A：周围的农田将因森林生长而变得更加茂密。B：附近的高速公路将被拆除，导致交通便利性降低。C：该综合体将占据更大的面积，以铺装基础设施取代部分农田。D：随着工业需求下降，该综合体将缩小。
正确答案： C

类别：未来预测（双图单轮）
图像分辨率： 10000x10000

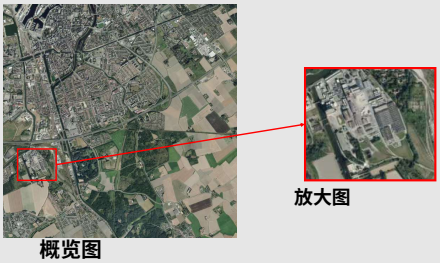
问题：
比较位于中央蜿蜒道路附近的农田图像 A 和图像 B，最有可能未来发生的变化是什么？根据观察到的视觉模式，该地区可能会发生什么变化？

选项：
A：田地可能会恢复为天然植被，因为可以看到作物
减少和杂草丛生的迹象出现在图像 A 和图像 B 之间。
B：这些田地很可能会被改建为住宅区，因为
从图像 A 到图像 B，沿道路的建筑物明显增加。
C：这些田地很可能会保持不变，因为没有明显的
迹象表明它们在图像 A 和图像 B 之间的结构或用途发生了变化。
D： 中央蜿蜒道路附近的农田很可能会
合并为更大的地块，因为一些较小的田地似乎已经
在图像 A 和图像 B 之间合并。
正确答案： D



图5. 不同任务的案例：异常检测（单图多轮）和未来预测（双图单轮）。

类别：未来预测（单图多轮）
图像分辨率： 10000x10000



问题1：
大型工业综合体位于图像中的什么位置?

选项：
A：图像的右上角。B：图像的左侧。
C：图像的左下角。D：图像的右侧。

正确答案： C

问题2：
什么关系或变化可能影响所选的工业综合体在不久的将来，根据图像中的视觉线索?

选项：
A：靠近主要公路，可能会提高该综合体的交通通达性。B：南侧的林地可能面临被清理以进行工业开发的风险。C：周围的大型住宅区可能会增加对工业服务的需求。D：附近存在农业用地，并可能向其扩张。

正确答案： D

问题3：
如果农业用地继续开发为工业区，所选工业综合体最有可能的未来状态是什么?

选项：
A：工业综合体的交通流量将增加，但由于空间限制，扩张幅度有限。B：由于分区法规，工业综合体将迁至离住宅区更远的地方。C：工业综合体将扩张，更多建筑物和基础设施将取代农田。D：工业综合体将保持不变，不会有明显变化。

正确答案： C

图6. 不同任务的案例：未来预测（单图多轮）。

类别：颜色检测 模型：Qwen2.5-VL-7B 提示词：你是遥感视觉问答（颜色检测）数据集的构建者。图像元数据：- 图像文件： {image path} - 分辨率： {width} x {height} - 参考边界框 [xmin, ymin, xmax, ymax]： [{xmin},{ ymin}, {xmax}, {ymax}] - 真实目标颜色： {target color}请生成一个严格符合以下要求的多项选择问答样本：1. 检查给定的边界框区域，识别主体对象。将“object”替换为简洁的对象类别名称（例如 car、truck、ship、building）。避免使用“thing”或“object”等含糊的词语。2. 完全照用以下问题模板（仅替换“object”占位词）：“确定图像中给定参考边界框内对象的颜色。图像分辨率： {width} x {height}。边界框： [{xmin}, {ymin}, {xmax}, {ymax}]。”3. 提供四个标记为 A、B、C、D 的答案选项：- 使用 {target color} （小写）作为唯一正确选项。- 选择三个不同的干扰颜色，这些颜色应合理地出现在图像的其他位置（其他实例或周围材料中）。使用简单的小写颜色词（red、blue、green、yellow、white、black、gray、brown、orange、purple）。- 确保四个选项各不相同，且正确颜色只出现一次。4. 将“answer”字段设为与 {target color}对应的选项字母（A/B/C/D）。5. 仅输出符合以下结构的 JSON：{ ”question”: ”...”, ”options”: {”A”: ”...”, ”B”: ”...”, ”C”: ”...”, ”D”: ”...”}, ”answer”: ”A|B|C|D”}

图7. 颜色检测的提示模板。

类别：形状/边缘识别 模型：Qwen2.5-VL-7B 提示词：你是一名遥感视觉问答（形状/边界描述）数据集构建者。图像元数据：- 图像文件： {image path} - 分辨率： {width} x {height} - 参考边界框 [xmin, ymin, xmax, ymax]： [{ xmin},{ymin}, {xmax}, {ymax}] - 真实目标形状： {target shape}请严格遵循以下要求，生成一个多项选择问答样本：1. 仅关注给定边界框内的主体对象；判断依据应是边界/形状，而非颜色或纹理。2. 完整照用以下问题模板（不要更改措辞，只替换数字）：“图像中给定参考边界框内主体对象的边界特征/形状，最准确的描述是什么？图像分辨率： {width}x {height}。边界框： [{xmin}, {ymin}, {xmax}, {ymax}]。”3. 提供四个答案选项，分别标为 A、B、C、D：- 仅将 {target shape} （小写）作为正确选项。- 编写三个各不相同的干扰项：可以是图像其他位置可见物体的合理形状描述，也可以是边界框内可能出现但不正确的印象（非主要元素）。- 选项应简洁，使用小写单词/短语，例如 rectangular、square、circular、oval、elongated、triangular、l-shaped、star-shaped、concave、convex、irregular、curved、linear、arc-shaped。- 确保四个选项各不相同，且正确形状仅出现一次。4. 将“answer”字段设置为对应于 {target shape}的选项键（A/B/C/D）。5. 仅输出 JSON，结构如下：{ ”question”: ”...”, ”options”: {” A”: ”...”, ”B”: ”...”, ”C”: ”...”, ”D”: ”...”}, ”answer”: ”A|B|C|D”}

图8. 形状/边缘识别的提示模板。

类别：方向检测 模型：GPT-5 thinking 提示词：你是一名专注于方向检测的遥感视觉问答数据集构建者。图像元数据：- 图像： {image path} - 分辨率： {width} x {height} - 参考边界框 [xmin, ymin, xmax, ymax]： [{xmin}, {ymin},{ xmax}, {ymax}]- 真实方向标签（必须是正确答案）： {target orientation} - 允许使用的方向词：{orientation vocab}你的任务是严格遵循以下规则，生成且仅生成一个选择题问答对：1. 检查边界框内的目标区域，识别主体对象（例如汽车、船舶、储罐、建筑物）。选择简洁、具体的对象类别名称。不要使用“物体”或“东西”等含糊词语。2. 使用以下确切的问题模板（仅替换占位内容）：“判断图像中给定参考边界框内 {{object name}} 的方向。图像分辨率：{ width} x {height}。边界框： [{xmin}, {ymin}, {xmax}, {ymax}]。”将 {{object name}} 替换为你选择的对象名称。3. 生成四个答案选项，分别标为 A、B、C、D。每个选项必须是从 {orientation vocab}中选出的一个方向词。- 将 { target orientation} 作为唯一正确选项。4. 将“answer”字段设置为对应于 {target orientation}的选项键（A/B/C/D）。5. 仅输出 JSON，结构如下：{ ”question”: ”...”, ”options”: {”A”: ”...”, ”B”: ”...”, ”C”: ”...”, ”D”: ”...”}, ” answer”: ”A|B|C|D”}

图9。方向检测提示模板。

类别：物体分类 模型：Qwen2.5-VL-7B 提示词：你是一名专注于物体分类的遥感视觉问答数据集构建者。图像元数据：- 图像文件： {image path} - 分辨率： {width} x {height} - 参考边界框 [xmin, ymin, xmax, ymax]:[{xmin}, {ymin}, {xmax}, {ymax}] - 真实目标类别： {target label} - 类别词表（有效类别名称）：{class vocab}严格遵守以下所有规则，生成且仅生成一组选择题问答对：1. 原样使用以下问题句子（数字已填入；不得更改措辞或标点）：“确定图像中给定参考边界框内主体对象的类别。图像分辨率： {width} x {height}。边界框： [{xmin}, {ymin}, { xmax}, {ymax}]。” 2. 选项 A/B/C/D 必须仅使用简洁的英语（每项最多 ≤ 6 个单词）：- 正确选项以标签 {target label} 开头。你最多可以在其前面添加两个形容词进行细化（例如，“dense {target label} zone”）。正确选项中必须且只能出现一次标签 {target label} 。- 从以下两个来源中创建三个不同的干扰项：(A) 同一图像中边界框以外、靠近边界框或特别显眼且具有视觉依据的物体；(B) 与目标属于同一上位类型、在语义上容易混淆的类别。至少一个干扰项必须来自 (A)，至少一个必须来自 (B)。若图像中没有明显可见的 (A) 类对象，则增加 (B) 类干扰项。- 干扰项必须使用包含 {class vocab} 中某一标签的英文描述词，但不得重复使用 {target label}。不得包含任何中文。各选项不得重复标签或相同措辞。3. “answer”字段必须恰好是指向包含经细化英文描述 {target label} 的选项的键（A/B/C/D）。4. 仅输出符合以下架构的 JSON：{ ”question”: ”...”, ”options”: {”A”: ”...”, ”B”: ”...”, ”C”: ”...”, ”D”: ”...”}, ”answer”: ” A|B|C|D”}

图 10。物体分类提示模板。

类别：相对空间关系模型：

Qwen2.5-VL-7B

提示词：你是一名专注于对象级相对空间关系的遥感视觉问答数据集构建者。你的目标：仅根据给定图像，生成一道以下类型的选择题：视觉问答-感知 → 对象级属性 → 相对空间关系。操作步骤（严格遵循）：1. 确定图像像素分辨率（$W \times H$）。2. 选择图像中两个相邻、显眼且清晰可见的物体或区域。• 物体应彼此不同、易于识别且常见（例如，颜色鲜艳或面积较大的区域，如体育场、森林、建筑物、河流、红色飞机）。• 每个物体都必须完整可见，不得被图像边缘裁切。• 两个物体之间必须有明确区分，且位置关系清晰。• 两个物体都应至少达到 10×10 像素，以确保清晰度。3. 按照以下确切句式，提出一个询问物体 A 相对于物体 B 的相对空间位置的问题："在图像中， <物体 A> 位于 <物体 B> 的什么位置？图像分辨率： {w} x {h}。" 详细描述每个物体的特征（例如，" 红色飞机 "" 正在航行的货船 "" 有 10 辆汽车的停车场"）。重要说明（必须遵循）：• 描述两个参照物体时，不要提及它们在图像中的全局位置（例如，"左上""右下"）。• 而应使用形状、颜色、纹理或类别等显著特征来描述它们（例如，"带红色跑道的圆形体育场""带白色屋顶的矩形灰色建筑"）。• 如果两个物体位于大致相同的区域（例如，二者都在右上方），不要透露这一点；模型应仅推断二者之间的相对空间关系。

• 如果物体很大，占据了图像的大部分区域，请避免重复提及绝对全局位置；应突出其独特的视觉特征。

• 通过方向、颜色或大小等差异，将每个物体与图像中其他相似物体区分开来（例如，可根据车位数量或附近的建筑物来区分多个停车场）。

• 问题应基于场景整体的环境语义，并涉及船只、房屋等较常见的物体。示例："在图像中，左侧有两架飞机的大型停车场相对于最大的机场航站楼位于何处？图像分辨率： {宽} × {高}。"答案选项：• 请恰好提供四个选项（A、B、C、D）。

• 选项必须是方向关系："上方"、"下方"、"左侧"、"右侧"、"左上角"、"右上角"、"左下角"、"右下角"。• 将选项顺序随机排列，并确保只有一个正确答案。质量要求：• 空间关系必须明确无歧义。• 不要选择相互重叠或含糊不清的物体。输出格式（仅限 JSON，不要添加额外文本）：{ "问题": "在图像中， <物体 A> 相对于 <物体 B>位于何处？图像分辨率： {宽} × {高}。"，"选项": {"A": "<方向>"，"B": "<方向>"，"C": "<方向>"，"D": "<方向>"}，"答案": "A|B|C|D" }约束：

• 仅输出 JSON 对象。• {宽} 和 {高} 必须为整数。• 答案必须恰好是 A/B/C/D 中的一个。

图11. 相对空间关系的提示模板。

类别：目标定位 模型：GPT-5 Thinking 提示词：你是一名专注于目标定位的遥感视觉问答数据集构建者。图像元数据：- 图像文件： {图像路径} - 分辨率： {宽度} x {高度}- 候选边界框（像素坐标，包含边界）：{候选边界框} - 真实目标边界框（必须是候选项之一）： {目标边界框}你的任务是创建且仅创建一个符合以下所有规则的选择题问答对：1. 原样使用此问题模板（仅替换为下方提供的物体短语占位符）："哪个边界框最准确地定位了 {物体短语}，分辨率： {宽度} x {高度}？" 2. 提供四个答案选项，标记为 A、B、C、D：• 每个选项值必须与上方列出的候选边界框字符串中的一个完全一致。• 每个候选边界框必须恰好使用一次（顺序不限）。3. 将"answer"字段设为与等于 {目标边界框}的候选边界框相对应的选项键（A/B/C/D）。4. 仅输出符合以下结构的 JSON：{ "question"："..."，"options"： {"A"："..."，"B"："..."，"C"："..."，"D"："..."}，"answer"："A|B|C|D"}

图12。目标定位提示模板。

类别：区域定位
模型：GPT-5 Thinking
提示词： 您是一名专注于区域定位的遥感视觉问答数据集构建者。
图像元数据：
- 图像文件： {图像路径} - 分辨率： {宽度} x {高度}- 候选边界框 (像素坐标，包含边界)：
{候选边界框} -真实目标边界框 （必须是候选框之一）： {目标边界框}
你的任务是创建恰好一个符合所有规则的选择题问答对：
1. 原样使用以下问题模板（根据下方提供的区域短语和候选框内的物体，
可以进行适当调整以替换占位符）：
"哪个边界框最能定位 {区域短语}，分辨率： {宽度} x {高度}？"
无论 {区域短语} 使用何种语言，均须将其转换为完整的英语并呈现在最终问题中。
2. 提供四个答案选项，分别标记为 A、B、C、D：
• 每个选项的值都必须与上面列出的候选边界框字符串之一完全匹配。
• 每个候选边界框都必须恰好使用一次（顺序不限）。
3. 将"answer"字段设为与 {目标边界框}相等的候选边界框所对应的选项键（A/B/C/D）。
4. 仅输出JSON，结构如下：
{
"question"："...",
"选项"：{"A": "...", "B": "...", "C": "...", "D": "..."},
"答案": "A|B|C|D"
}

图13. 区域定位的提示模板。

类别：物体计数 模型：
GPT-5 thinking
提示词：你是遥感视觉问答数据集的构建者。你的目标：仅根据给定图像，创建一道类型为：视觉问答-感知 → 计数与测量 → 物体计数的选择题。严格遵循以下步骤：1. 选择一个符合以下条件的物体类别：- 清晰可见且可逐个计数；- 尺寸不能太小（至少 10×10 像素），没有严重遮挡，且类别明确；- 用精确的属性描述（例如 “ 大型货船 ”“ 红色椭圆形跑道”“停车场里的白色小汽车”）；- 不要描述得含糊或过于宽泛（避免使用“船”“车辆”“建筑物”等标签，而应使用具体且有辨识度的描述，例如“大型油轮”“蓝顶两层建筑”）。如果存在非常小或难以辨认的物体，请忽略它们，仅选择醒目且易于识别的物体。2. 指定计数范围：- 如果物体分布在整张图像中，询问图像中的总数；- 如果物体集中在某个明确的区域（例如右下方），则询问该区域内的数量。区域描述必须准确、自然（例如“右下角靠近海岸线的区域”）。3. 准确统计物体数量。4. 按以下格式构造问题：“图像中 <详细物体描述> 有多少 <个 / [特定区域]>内有多少个？”5. 提供四个答案选项（A、B、C、D）：- 每个选项必须是整数；- 选项数值应接近正确数量（与正确数量相差 ±1–3）；- 随机安排正确答案的位置；- 避免使用 10/20/30/40 这类差距过大的简单选项。输出格式（仅限 JSON，不要添加额外文本）：{ ”问题”：”图像中有多少 <个>物体.....？ ”，”选项”： {”A”： <整数>，”B”： <整数>，”C”： <整数>，” D”： <整数>}，”答案”：”A|B|C|D”}

图14. 物体计数的提示模板。

类别：区域计数 模型：Qwen2.5-VL-7B 提示词：你是遥感视觉问答数据集的构建者。你的目标：仅根据给定图像，创建一道类型为：视觉问答-感知 → 计数与测量 → 区域计数的选择题。严格遵循以下步骤：1. 选择一个物体类别，要求该类别：- 清晰可见且可逐个计数；- 尺寸不能太小（至少 10×10 像素），没有严重遮挡，且含义明确；- 用精确的属性描述（例如“大型货船”“红色椭圆形跑道”“停车场里的小型白色汽车”）；- 描述不能含糊或过于宽泛（避免使用“船”“车辆”“建筑物”）。2. 指定计数范围：- 如果物体分布在整幅图像中，则询问整幅图像中的总数；- 如果物体集中在一个明显的区域（例如右下），则询问该区域中的数量。区域描述必须准确、自然（例如“海岸线附近的右下角”）。3. 精确计数。4. 按以下格式构造问题：“ <详细物体描述> 有多少个 <在整幅图像中 / 在 [特定区域内]>？”5. 提供四个答案选项（A、B、C、D）：- 每个选项必须为整数；- 选项的数值应接近正确数量（与正确数量相差不超过 ±1–3）；- 随机排列正确答案的位置；- 避免使用10/20/30/40这类差距过大的简单选项。输出格式（仅 JSON，不要添加其他文本）：{ ”question”：”有多少 <物体>.....？ ”，”options”： {”A”： <整数>，”B”： <整数>，”C”： <整数>，”D”： <整数>}，”answer”：“A|B|C|D”}

图15. 区域计数的提示模板。

类别：多区域联合对比（三图对比）模型：GPT-5 thinking

提示词：你是一个用于多区域联合对比的遥感视觉问答数据集构建者。任务：- 给你三张卫星图像（A、B、C），每张图像包含一个参考边界框和一个对比属性标签。- 仅根据给定属性比较三个参考区域（A/B/C）。- 只生成一道选择题，并使用固定选项。- 正确答案必须是 A、B 或 C。图像 A/B/C：- 文件：{图像路径 A/B/C}- 分辨率：{宽度 A/B/C}x {高度 A/B/C}- 参考边界框 [xmin、ymin、xmax、ymax]：[{xmin A/B/C}，{ymin A/B/C}，{xmax A/B/C}，{ymax A/B/C}] - 对比属性标签：“{标签 A/B/C}”问题要求：1. 问题应询问这三张图像中的哪张显示了上述标签所指属性的极值。例如：“这三张图像中的三个给定边界框，哪个建筑密度最高？分辨率：{宽度 A} x {高度 A}。图像 A 的边界框：[{xmin A}，{ymin A}，{xmax A}，{ymax A}]。图像 B 的边界框：[{xmin B}，{ymin B}，{xmax B}，{ymax B}]。图像 C 的边界框：xmin C，ymin C，xmax C，ymax C。”此处的“最高建筑密度”应根据标签所隐含的属性确定。- 如果标签没有明确说明最高或最低，则选择最自然的极值（例如“最高”）。- 问题应简洁（不超过25个词）。- 三张图像的分辨率相同，因此只需提及一次。- 无论标签使用何种语言，都要将其转换为英语问题。2. 选项必须为固定字符串：- A：“图像 A”- B：“图像 B”- C：“图像 C”- D：“无法比较”3. 根据检查三张图像的结果，将“answer”设为唯一符合正确极值的选项键（A/B/C/D）。输出格式（仅 JSON，ASCII 字符，不要添加说明）：{ ”question”：”……”，”options”：{”A”：“图像 A”，“B”：“图像 B”，“C”：“图像 C”，“D”：“无法比较”}，”answer”：“A|B|C”}

图16. 多区域联合对比的提示模板。

类别：多区域联合对比（单张图像）模型：GPT-5 Thinking 提示词：你是一名遥感视觉问答数据集构建者，专注于多区域联合对比（单张图像）。图像元数据：- 图像文件：{图像路径} - 分辨率：{宽度} x {高度} - XML 中标注的对比区域：{候选区域}以上每行的格式为：[xmin, ymin, xmax, ymax] — “区域描述” 你的任务：1. 仔细阅读区域描述。描述中已明确指出对比属性（例如建筑密度、交通流量）。2. 撰写一道且仅一道比较这些区域的选择题。问题必须：- 严格使用以下模板（仅将大括号内容替换为你的措辞）：“哪个边界框最能代表 {{比较短语}}” - 确保比较短语忠实于所提供的描述（例如“建筑密度最高的区域”、“绿度最低的农田地块”）；- 如果描述暗示极值（最高/最低/最多/最少），则据此提问；否则采用自然的比较措辞（例如“规模最大的紧凑型住宅群”）；- 问题简洁，不超过 25 个词。3. 选项 A、B、C、D：- 完全照抄所提供的边界框字符串，不得改动。不得编造新的边界框；- 每个标注框必须在选项中恰好出现一次；- 如果标注框少于四个，则用字符串“-”（单个连字符）填充剩余选项。4. 根据描述的属性，通过目视比较图像中的区域，确定正确答案。- 将“answer”设为正确边界框对应的选项键（A/B/C/D），绝不能设为“-”。5. 仅输出 JSON，格式如下（不添加说明，仅使用英语）：{ ”question”: ”...”, ”options”: {”A”: ”...”, ”B”: ”...”, ”C”: ”...” }, ”answer”: ”A|B|C|D”

图17. 提示词：多区域联合对比（单张图像）。

类别：目标状态判断（多轮对话）模型：GPT-5

Thinking

提示词： 你是一名遥感视觉问答数据集构建者。
任务：设计一段三轮对话（用户 ↔ 助手），围绕图像中一个显著目标的**目标状态判断** 展开。
严格要求：
1. 仔细检查图像，选择一个其状态可根据可见证据明确判断的目标。
2. 描述该目标的显著特征，使其不会与附近的相似物体混淆。
3. 编写三轮互动：- 每轮都是一道包含四个选项（标记为 A/B/C/D）的选择题，由用户提出。- 每个选项都必须以图像为依据。根据其他视觉线索创建三个合理的干扰项，另加一个正确答案。- 助手只能回复正确选项的字母，后面紧跟选项原文（格式："<字母>. <选项文本>"）。不要输出解释、推理或任何额外文字。- 第 1 轮：询问能 够揭示物体状况的视觉线索。四个选项必须明显不同，且不能表达相同的含义。在问题中指出物体在图像中的大致位置，并详细描述该物体，但不要包含任何暗示答案的线索。- 正确选项必须依赖需要观察图像才能获取的信息，而不能仅凭常识。- 第 2 轮：询问这些线索如何影响功能或运行状态。四个选项必须显然彼此不同，且不得表达相同含义。- 第 3 轮：询问状态分类；四个选项可以从以下固定列表中的允许的状态中选取。也可以根据所选物体的常识属性添加其他选项，但这些选项必须与正确答案及彼此不同。
允许的状态：”移动中”、”静止”、”使用中／活跃”、”未使用／非活跃”、”完好”、”受损”、”正在建-设”、”有人居住”、”空置”、”被雪覆盖”、”被淹”、”干燥”、”有植被覆盖”、”起飞”、”着陆”、”停放”、”航行”、”飞-靠岸”，”倒车”，”停止”，”运输途中”，”拆除中”，”翻修中”，”空置”，”巡航”，”滑行”，”已离开”，”已抵达”，”靠岸”，”系泊”，”在干船坞中”，”漂流”，”遭侵蚀”，”烧毁”，”泥泞”，”已灌溉”，”冻结”，”杂草丛生”，”干旱”，”倒塌”，”不稳定”，”受保护”，”无法进入”，”严重损坏”，”部分损坏”，”完全正常运行”，”非正常运行”
输出格式（仅限 JSON）： {
”目标物体”: {
”description”: ”<区别性描述>”,
”状态证据”: [”<关键线索 1>”, ”<关键线索 2>”]
},
”对话”: [
{
”轮次”: 1 ,
”用户问题”: ”<A-D问题>”,
”选项”: {”A”: ”<选项>”, ”B”: ”<选项>”, ”C”: ”<选项>”, ”D”: ”<选项>”},
”助手回答”: ”<字母>. <选项文本>”
},
{
”轮次”: 2 ,
”用户问题”: ”<A-D 题目>”,
”选项”: {”A”: ”<选项>”, ”B”: ”<选项>”, ”C”: ”<选项>”, ”D”: ”<选项>”},
”助手回答”: ”<字母>. <选项文本>”
},
{
”轮次”: 3 ,
”用户问题”: ”<使用允许的状态作答的 A-D 题目>”,
”选项”: {”A”: ”<状态>”, ”B”: ”<状态>”, ”C”: ”<状态>”, ”D”: ”<状态>”},
”助手回答”: ”<字母>. <状态文本>”
}
]
}

图18. 目标状态判断（多轮对话）的提示模板。

类别：目标状态判断（单轮）模型：GPT-5 thinking

提示词：你是一名专注于目标状态判断的遥感视觉问答数据集构建者。图像元数据：- 图像文件： {image path} - 分辨率： {width} x {height} - 允许使用的状态词汇（请严格使用以下字符串）： {state vocab}你的任务是严格按照以下要求创建且仅创建一个选择题问答对：1. 仔细检查图像，选择一个状态在视觉上明确无误的显著目标。- 描述该目标的独特外观（大小、形状、颜色、背景），确保不会与相邻目标混淆。- 指出该目标在图像中的大致位置（例如“左上方”“中下部区域”）。2. 严格使用以下模板撰写问题（仅替换占位符）：“图像中位于 {{object description}} 的 {{location phrase}} 的目标当前处于什么状态？”- 将{object description} 替换为对目标的详细描述（确保详细描述不包含可用于推断正确答案的提示）。- 将 {location phrase} 替换为位置短语（例如“图片右上角”）。- 问题文本中不得泄露正确答案。3. 提供四个答案选项，分别标记为 A、B、C、D。正确状态须恰好出现一次；其余三个选项必须是合理的干扰项。4. 将“answer”字段设为对应正确状态的选项键（A/B/C/D）。5. 填写“target object”字段：- “description”：第 1 步中的目标描述。- “location”：第 1 步中的位置短语。- “state”：正确状态（必须与答案选项文本完全一致）。- “evidence”：包含两条简洁视觉线索的列表，用于支持状态判断。6. 仅输出以下结构的 JSON（不得添加额外文字、解释或 Markdown）：{ ”target object”： { ”description”：”...”，“location”：”...”，“state”：”...”，“evidence”： [”...”，“...”] }，“question”：“...”，“options”： {”A”：“...”，“B”：“...”，“C”：“...”，“D”：“...” }，“answer”：“A|B|C|D”}

图19. 对象状态判断提示模板（单轮）。

类别：异常检测与解读 模型：GPT-5 Thinking

提示词： 你是一名遥感视觉问答助手，负责对图像进行异常检测与解读。规则：1. 仔细观察图像，并选择一个状态明确的物体。- 使用具有区分性的特征描述该物体（例如，大小、形状、颜色、位置）。- 提供该物体在图像中的大致地点（例如“左上方”）。2. 在你的问题中，根据所选物体的视觉特征，包括其在图像中的位置，对其进行详细描述。- 示例：“停放在河岸边的红色货船，图像右侧可以看到一座蓝色仓库。这艘货船位于图像中的什么位置？”- 确保问题和描述与图像中的视觉元素明确相关。- 问题不应包含任何答案提示。3. 随后，助手应根据图像中的可见线索，询问导致该物体当前状态或异常的最可能的原因。例如：“货船停放在河岸边最可能的原因是什么？”4. 提供四个选项（A、B、C、D），其中：- 一个选项应是根据图像视觉证据得出的最合理原因；- 其他三个选项应是合理但不正确的干扰项。5. 确保问题的答案取决于图像中的具体视觉线索。答案不得依赖常识或图像中未显示的外部因素。输出格式（仅限 JSON，不得添加其他文本）：{”question”：”<包含物体描述和地点的详细问题>”，”options”： { ”A”：”...”，”B”：”...”，”C”：”...”，”D”：”...” }，”answer”：”A|B|C|D”}

图20. 异常检测提示模板。

类别：异常检测与解读（三轮对话）模型：GPT-5 Thinking

提示词：[illegible]区域，设计一段关于异常检测与解读的三轮对话（用户 ↔ 助手）。要求：1. 检查图像，并选择一个适合进行未来预测的目标或区域，其视觉状态和周边环境清晰可辨。- 所选目标或区域应清晰可辨，且位于图像中的独特位置。- 尽量选择相对独特、而非图中有许多相似对象的目标。- 尽量选择在当前分辨率下能够详细观察、而在较低分辨率下无法辨别其细节的目标。2. 第一轮（定位）。详细描述目标或区域，包括其区别性特征和大致位置（例如左上方、右侧）。- 描述必须确保目标独一无二，不能同时指代图像中的多个目标。- 问题应为：“图像中 [目标/区域] 位于哪里？”（使用此措辞）。- 第一轮问题中对 [目标/区域] 的描述必须极为详细，以便与其他相似目标区分开来。- 描述应包含目标的特征，并指出其与周边目标之间的关系，使描述更加详尽。- 描述必须超过五个词且少于 50 个词，包含目标固有特征（仅限描述颜色、形状、大小和方向；不得描述状态）以及与周边目标的空间关系。- 避免“货船”这类过于简单的描述。- 答案选项应为目标/区域在图像中的位置。选项只能描述其在图像中的位置，不得提及其他内容，例如与其他目标的关系。示例：“图像左侧。”避免使用类似“图像左侧，靠近河岸”这样的答案。3. 第二轮（状态分析）。分析已识别目标的状态。根据图像中的可见证据，考虑该目标可能处于的状态或状况。本轮重点是了解目标当前的状况或状态，以帮助判断下一轮中的可能原因。- 问题示例：“图像中 [目标/区域] 当前的状态或状况可能是什么原因造成的？”- 选项示例：A. 目标处于常见状况，未观察到异常。B. 目标似乎因天气或环境等外部因素而静止。C. 目标处于异常状态，可能由机械故障或环境变化造成。D. 目标属于常规作业的一部分，没有显著变化。- 本轮尚不提供最终原因，而是旨在根据视觉线索缩小可能情形的范围。4. 第三轮（原因判定）。基于前述分析，判断目标当前状态或状况最可能的原因。结合图像中的视觉证据和前几轮的推理，给出答案。- 问题：“鉴于目标当前的状态，这种状况最可能是什么原因造成的？”- 选项：A. [基于可见证据最合理的原因]。B. [合理的干扰项 1]。C. [合理的干扰项 2]。D. [合理的干扰项 3]。- 本轮应侧重于根据先前观察到的视觉证据，给出最可能的原因，而非依赖外部知识。- 答案应基于前几轮中的视觉线索和状态分析，而非外部知识。助手只需回复正确的字母和选项文本（“< 字母 >。 < 选项文本>”）。输出格式（仅 JSON，不附加其他文本）：{”问题 1”: {”用户问题”: ”图像中 [目标/区域] 位于哪里？”, ”选项”: { ”A”: ”...”, ”B”: ”...”, ”C”: ”...”, ”D”: ”...” }, ”正确答案”: ”<字母>。 <正确选项文本>”}, ”问题 2”: {”用户问题”: ”什么关系或变化可能会在不久的将来影响所选目标/区域？”, ”选项”: { ”A”: ”...”, ”B”: ”...”, ”C”: ”...”, ”D”: ”...” }, ”正确答案”: ”<字母>。 <正确选项文本>”}, ”问题 3”: {”用户问题”: ”如果 [第二轮的答案] 发生，所选目标/区域最可能呈现什么未来状态？”, ”选项”: { ”A”: ”...”, ”B”: ”...”, ”C”: ”...”, ”D”: ”...” }, ”正确答案”: ”<字母>。 <正确选项文本>”}}

图21 . 异常检测（多轮对话）提示模板 .

类别：未来预测（三轮对话）模型：GPT-5 thinking

提示词：你是遥感视觉问答数据集构建者。任务：围绕卫星图像中的一个显著目标或区域，设计一段关于未来预测的三轮对话（用户 ↔ 助手）。要求：1. 检查图像，并选择一个适合进行可追溯性分析的目标或区域，其视觉状态和周边环境清晰可辨。- 所选目标或区域应清晰可辨，且位于图像中的独特位置。- 优先选择相对独特、而非图中有许多相似对象的目标。优先选择在当前分辨率下能够充分观察、而在较低分辨率下无法辨认其细节的目标。2. 第一轮（定位）：- 详细描述目标或区域，包括其区别性特征和大致位置（例如左上方、右侧）。- 描述必须确保目标独一无二，不能同时指代图像中的多个目标。- 问题应为：“图像中 [目标/区域] 位于哪里？”（使用此措辞）。- 第一轮问题中对 [目标/区域] 的描述必须极为详细，以便与其他相似目标区分开来。- 描述必须：1.包含目标特征，并指出其与周边目标的关系，使描述更加详尽；2.超过 5 个词且少于 50 个词；3.包含目标的固有特征（仅限描述颜色、形状、大小和方向；不得描述状态）；4.包含与周边目标的空间关系；5.避免“货船”这类过于简单的描述。3. 答案选项应为目标/区域在图像中的位置，只描述位置（不得提及与其他目标的关系等其他内容）。示例：“图像左侧。”4. 第二轮（关系/变化）：- 分析已识别目标或区域的状态。根据可见证据，考虑可能在不久的将来影响它的关系或变化。- 本轮重点是根据图像中的视觉线索，了解可能影响所选目标或区域的关系或变化。- 问题示例：“根据图像中的视觉线索，什么关系或变化可能会在不久的将来影响所选目标/区域？”- 选项应为四个彼此不同、互不重叠的可能关系或变化描述（A、B、C、D）。5. 第三轮（未来状态）：- 基于前述分析，判断如果第二轮中的关系或变化发生，所选目标或区域最可能呈现的未来状态。- 问题示例：“如果 [第二轮的正确答案] 发生，所选目标/区域最可能呈现什么未来状态？”- 选项：A：可能的未来状态 A；B：可能的未来状态 B；C：可能的未来状态 C；D：可能的未来状态 D。- 正确选项应基于可见证据和前几轮的推理，选出最合理的未来状态，而非依赖外部知识。输出格式（仅 SON，不附加其他文本）：{”问题 1”: {”用户问题”: ”图像中 [目标/区域] 位于哪里？”, ”选项”: { ”A”: ”...”, ”B”: ”...”, ”C”: ”...”, ”D”: ”...” }, ”正确答案”: ”<字母>。 <正确选项文本>”}, ”问题 2”: {”用户问题”: ”什么关系或变化可能会在不久的将来影响所选目标/区域？”, ”选项”: { ”A”: ”...”, ”B”: ”...”, ”C”: ”...”, ”D”: ”...” }, ” 正确答案 ”: ”< 字母 >。 < 正确选项文本 >”}, ” 问题 3”: {” 用户问题 ”: ” 如果 [第二轮的答案] 发生，所选目标/区域最可能呈现什么未来状态？”, ”选项”: { ”A”: ”...”, ”B”: ”...”, ”C”: ”...”, ”D”: ”...” }, ”正确答案”: ”<字母>。 <正确选项文本>”}}

图22. 未来预测（多轮对话）的提示模板。

类别：未来预测（双图像时间对比）模型：GPT-5 Thinking

提示词： 你是遥感视觉问答数据集的研究

提供了同一地点在不同时间拍摄的两张卫星图像 （图像 A = 较早年份，图像 B = 较晚年份）。请设计一道选择题，要求通过比较这两张图像来预测物体的未来状态。要求：1. 仔细观察两张图像，并找出一个在图像 A 和图像 B 中都清晰可见的显著目标或区域。• 选择一个随时间发生的变化在视觉上明显且细微的物体或区域。不要依赖广为人知的事实；变化必须能从图像本身清楚地观察到。• 避免选择频繁出现或无需仔细查看图像就能轻易识别的物体。2. 编写一个问题，要求：

• 通过其作用或大致位置来指明所选物体或区域，明确提及“图像 A”和“图像 B”，但不要明确描述诸如颜色、形状或纹理等视觉特征；

• 询问根据两幅图像之间观察到的视觉变化，物体最有可能呈现的未来状态；• 指示回答者依据两幅图像中的视觉线索作答，而非依赖外部知识。3. 提供四个答案选项（A、B、C、D）：

• 一个选项必须描述最有可能的未来结果，且以两幅图像中观察到的视觉证据为依据；
• 另外三个选项应是合理但不正确的干扰项，聚焦于不受视觉证据支持的未来状态或趋势；

• 保持选项简洁，重点关注未来预测，同时融入支持该选择的视觉观察；• 正确答案应结合两幅图像中的视觉线索，并解释未来预测。例如：“道路附近绿地的减少以及居民点的增加表明，这些居民点可能会进一步扩张。” 4. 确保答案取决于对图像 A 和图像 B 的对比。问题必须只能通过直接比较两幅图像来回答，并依靠细微的视觉证据推断未来预测。5. 输出必须是有效的 仅限 JSON：{”问题”：”<包含两幅图像视觉变化的未来预测问题>”，”选项”： {”A”：”<基于视觉线索的未来状态选项>”，”B”：”<未来状态选项>”，”C”：”<未来状态选项>”，”D”：”<未来状态选项>”}，”答案”：”A|B|C|D”}

图23. 未来预测（双图时间对比）的提示模板。

类别：答案评估（开放式视觉问答，无选项）模型：GPT-5 Thinking

提示词： 你是一名开放式视觉问答（无选项）领域的专家评审员。将参考答案视为基于图像得出的真实标注。避免奖励缺乏依据的具体细节；但是，对于轻微且不矛盾的补充说明，可以接受，不应过于严厉地扣分。

根据以下标准打分，分值范围为1–100：1. 与参考答案的正确性／一致性（最重要）。2. 实用性／完整性：直接回答问题，并在相关时包含必要属性／单位。3. 幻觉／捏造：仅当细节缺乏依据且会实质影响答案时才扣分；无害的额外背景信息不影响评分。宽松匹配准则：• 语义等价视为正确：同义词、常见改述、相近的颜色类别（例如，青绿色 ≈ 蓝绿色；栗色≈ 深红色），以及合理的方向映射（上 ≈ 上方；左 ≈ 当与图像坐标一致时可映射为西方）。

• 数值／计数容差（宽松）：如果参考值是整数 n，则允许 ± 1 当 $n \leq 10$；当 $n > 10$，允许约$\pm 20\%$ 的舍入误差。• 范围／近似值：落在参考值附近合理范围内的答案均可接受。

• 单位：正确换算或明显/隐含的单位均可接受；缺少单位只作轻微扣分，除非会改变含义。• 是非题：答案必须与参考答案一致。带有保留措辞但正确的回答（例如“可能是”）也可接受。• 如果参考答案为“未知/无法确定/无法从图像中推断”，优先接受承认不确定性的回答。• 不要过度惩罚风格或语法；重点关注事实内容及其与参考答案的一致性。建议的（非强制）分数区间：• 80–100：语义等同且准确；没有实质性幻觉；包含所需信息。• 70–80：基本正确；有轻微遗漏或细微偏差；没有矛盾。• 60–70：部分正确；包含一些关键信息或接近的数值；存在轻微问题。• 30–60：大部分错误、与参考答案矛盾，或包含明显/有意义的幻觉。• 0：空置、离题或内容不可接受。问题： {q}参考答案（视为图像真实标注）： {ref}模型回答： {ans}仅返回两行：score: <1-100>

图24. 开放式视觉问答答案评估的提示模板（评分范围为1–100）。

类别：全球场景描述（单幅航空/卫星图像）模型：GPT-5 Thinking

提示词覆盖整个面积区域、包含城市、乡村、工业、交通和自然要素的分辨率航空/卫星图像，按照以下规则生成详细的场景描述。

1. 总体描述： 用 一个句子 概括主要土地利用与地貌的组合、空间布局和主要功能。说明城市区域、农田、工业区、交通走廊、港口和绿地等关键区域的 相对位置及其作用（例如沿海布局、沿河平行分布的带状区域，或周围环绕农田的集聚型城市核心区）。

2. 子区域描述： 描述左上、右上、左下和右下象限。针对每个象限，用 1–3 个句子 说明主要物体、其空间组织方式以及主要功能或用途（例如灌溉农业、港口物流、高密度居住核心区、工业与居住混合区）。明确说明各区域之间的关系，例如“道路沿海岸线延伸”“住宅街区与田地之间隔着一条高速公路”或“工厂聚集在河流旁”。使用“多家工厂”“一条主要高速公路”或“多个住宅小区”等具体表述，避免使用“很多”或“一些”等含糊词语。

3. 详细描述要求： 物体分类： 涵盖 自然要素 景观（河流、山脉、湖泊、森林、湿地）以及 人工要素结构（道路、桥梁、铁路、机场、港口、工业园区、住宅区、商业区）。空间关系： 清晰描述物体的排列方式，例如道路沿海岸线延伸、建筑物与铁路平行、街道和街区呈网格状交叉、房屋散布在丘陵地形上，或工厂沿滨水区聚集。 功能/用途： 根据可见形态推断各区域的用途，例如农作物种植农田、结合住宅和商业建筑的城市中心、用于生产和仓储的工业区，或用作货物配送枢纽的港口。相比模糊的数量词，优先使用“多座大型工厂和数个集装箱堆场”或“分布稀疏的住宅区”等具体表述。

4. 输出语言： 使用清晰、简洁的 英语提供描述，重点关注关键要素、其空间布局及相互关系。避免过于笼统的措辞，描述场景中有视觉依据的内容及其推断出的功能。

输出示例（供参考）：
整体描述： 一座沿海港口城市占据中央区域，南部海岸线上有一座繁忙的深水码头，一条高速公路沿海岸延伸，连接港口与内陆工业带；周边住宅和商业区被北部和西部的农田及天然植被环绕。

子区域描述： 在左上象限，矩形农田占据主导，笔直的灌溉渠大致沿南北方向延伸，构成田地的边界；田地边缘散布着农舍，表明这里进行着集约化作物种植。右上象限分布着一个密集的住宅和商业区，街道布局呈网格状，其中点缀着数座公园。一条大河穿过该区域，将城市核心区与港口连接起来。左下象限是一个工业区，大型工厂沿平行的行列布局，并通过道路和铁路网络与滨水区相连；仓库分布在园区外围。右下象限可见沿蜿蜒的地方道路分布的郊区住宅群，其间点缀着绿地和小型公园；一条狭窄的河流蜿蜒穿过该区域，并在港口附近汇入更大的水体。

图25。提示词：用于航空/卫星图像的全局场景与子区域描述。

类别：结构化描述（基于象限的场景描述）模型：GPT-5 thinking

提示词： 你是一位专业的遥感分析师　　. 请为每幅高分辨率航空/卫星图像生成结构化描述并严格遵循以下模板（不要添加额外章节）：

总体概述:
- [一句话描述主要土地利用/地貌组合、空间布局和主要功能。]

左上象限:
- [用两到三句话详细描述主要自然要素和人工要素、它们的空间组织、明确的关系（例如，“运河环绕矩形稻田”，以及推断出的用途。]

右上象限:
- [此象限的细节程度相同。]

左下象限:
- [该象限的细节程度相同。]

右下象限:
- [该象限的细节程度相同。]

对象 + 用途分类:
-自然要素: [列出河流、森林、丘陵、湿地、海岸线等，并说明其作用。]
-人工要素: [列出道路、铁路、桥梁、港口、工厂、仓库、住宅类型等，并说明其作用。]

指南:
1. 提及自然景观和建成基础设施，强调空间关系（沿河平行延伸的道路、码头附近的仓库簇、由树带隔开的农田等）。2. 对每个象限，说明其主要功能或用途（例如商业物流、灌溉农业、高密度居住核心区、工业与居住混合区）。3. 避免使用模糊的量词，例如“一些”或“许多”；应使用具体描述，如“5座狭长仓库”“密集的低层住宅网格”或“蜿蜒向东西延伸的狭窄河流”。4. 语言必须为英语，简洁而具体，并在适当时说明物体之间及各象限之间的相对位置。5. 切勿包含占位文本、道歉内容，或提及“提示词”或“图像”；只关注可观察到的证据和推断出的用途。

图26。提示词：针对高分辨率遥感影像的基于象限的结构化描述。